BIOINFORMATIC TOOLS FOR TESTING MICROBIAL ECOLOGY THEORY IN NATURAL ENVIRONMENTS THROUGH METAGENOMICS

A Dissertation
Presented to
The Academic Faculty

By

Luis Miguel Rodriguez Rojas

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics
School of Biological Sciences

Georgia Institute of Technology

December 2016
Copyright © 2016 by Luis Miguel Rodriguez Rojas

BIOINFORMATIC TOOLS FOR TESTING MICROBIAL ECOLOGY THEORY IN NATURAL ENVIRONMENTS THROUGH METAGENOMICS

Approved by:

Dr. Konstantinos T. Konstantinidis, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Mark Borodovsky
School of Biomedical Engineering
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Joel Kostka
School of Biological Sciences
Georgia Institute of Technology

Dr. James M. Tiedje
NSF Center for Microbial Ecology
Michigan State University

Date Approved: November 11, 2016

To Nelson and M[a] Eugenia Rodríguez Rodríguez,
Tomás Leonardo Aguilar Rojas,
and Martín Álvarez Flautero.
Remote sources of peace and inspiration.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Konstantinidis, whose continuous support, motivation, and inspiration made the current dissertation possible and the process leading to it a most enriching and enjoyable experience. From his many virtues as advisor, I express my most sincere appreciation for his relentless questioning of my ideas and continuous encouragement to question his own; routinely begetting free and open scholastic discussion.

I thank my committee members, Dr. Borodovsky, Dr. Jordan, Dr. Kostka, and Dr. Tiedje, whose much appreciated advice was always timely and insightful. My sincere thanks to Luis (Coto) Orellana, Despina Tsementzi, Chengwei Luo, and all the members of the Kostas Lab at Georgia Tech; to Mehmet Belgin and all the team of the Partnership for an Advanced Computing Environment (PACE) at Georgia Tech; to Will Overholt, Dr. Joel Kostka, and Dr. Chris Gaby from the Joel Kostka Laboratory at Georgia Tech; and to Dr. James R. Cole, Santosh Gunturu, Dr. James M. Tiedje, and all of the RDP team at Michigan State University. My collaborations with all of them made possible this dissertation and I greatly appreciate the help and advice I received from each one of them.

I thank my beloved friends Despina Tsementzi, Suehayl Sadik, Natasha de León, and Eleni Vaiopoulou for their companionship and friendship, for the many memories we built together, and for the numerous smiles and laughter along the recent years. I would like to thank my friends in the distance, especially Camilo Alejo Monroy, Juliana Gil Urbano, Laura Perlaza, Sergio Andrés Álvarez, Lorena Margarita Flautero, Natalia Forero Serna, Laura Catalina Herrera, Margie Álvarez, and Andrés Jiménez, with whom even infrequent contact would make

me frequently happy. I would also like to thank Hannah Nicol, Jack Lin, and Faye Jonah, whose company and dances made the past months unforgettable. Finally, I would like to thank my parents Maria Elena and Luis Miguel, my siblings Luisa Fernanda, Miguel Ángel, and Tomás, my grandparents Jesús Antonio, Ana Leonor, Margarita, and Hilda, my uncles, aunts, cousins, and all of my family for their love, encouragement, and inspiration.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The study of microbial ecology has been traditionally hampered by the inability to sample members of microbial communities uniformly at random in their natural environments. However, advances in molecular techniques during the past three decades have allowed the characterization of communities through DNA census. The existence of a global phylogenetic reference framework (Woese & Fox, 1977) sparked the popularization of 16S/18S ribosomal RNA gene amplification (SSU-rRNA amplicons) for the characterization of microbial communities. The use of SSU-rRNA amplicons has been further promoted by the availability of large standardized reference databases such as Ribosomal Database Project – RDP (Cole et al., 2014), allowing unprecedented advances in microbial ecology. However, the SSU-rRNA universality implies a degree of conservation that comes at the expense of low resolution near and below the species level (Cole, Konstantinidis, Farris, & Tiedje, 2010; Rodriguez-R, Castro, & Konstantinidis, In preparation). In order to solve this shortcoming for isolated organisms, recent advances in whole-genome comparisons have provided the framework necessary to re-define the bacterial and archaeal species on the basis of genome-aggregate Average Nucleotide Identity –ANI– (Konstantinidis, Ramette, & Tiedje, 2006; Konstantinidis & Tiedje, 2005a, 2005b). The extension and application of this theoretical framework to the study of natural populations and communities is now possible thanks to the availability and increasing popularization of metagenomics. Such advance has the potential to bring species-level resolution to the characterization of microbial communities, and is the subject of **chapter I**. However, the feasibility of such application is fully realized only when the proper tools and techniques are made available. Therefore, a guide to computational tools to explore and quantitatively compare

metagenomic datasets is presented in **chapter II**, and a suite of bioinformatic tools for genomics and metagenomics, the enveomics collection, is presented in **chapter III**. A particularly pervasive but underappreciated problem in the use of metagenomics is the issue of sequencing coverage, *i.e.*, the fraction of the microbial community in a sample characterized by sequencing, potentially decreasing the accuracy of both individual sample characterizations and comparative analyses, as discussed in **chapter IV**. In order to accurately assess sequencing coverage we developed Nonpareil, a computational tool that accurately estimates abundance-weighted average coverage in a metagenomic sample using read redundancy, as described in **chapter V**. Moreover, Nonpareil can be used to determine the sequence diversity in a community independently of databases or sequence coverage, allowing higher accuracy in the determination of alpha-diversity, as described in **chapter VI**. Chapter VI also describes recent computational optimizations on Nonpareil 3 using *k*-mer matching and high-performance computing in order to cope with the increasing volume of data that becomes available from environmental or clinical metagenomics surveys. The application of novel computational and statistical techniques, including those presented here, have the potential to close the gap between ecology theory and testing in microbial systems. We addressed factors that drive microbial community assembly using time-series metagenomics in two different ecosystems. First, we documented the post-disturbance successional patterns in shoreline sediments in Pensacola beach (Florida, USA) after the large-scale deposition of hydrocarbons caused by the 2010 Macondo oil spill in the Gulf of Mexico. This study and our approach to test the specialization-disturbance hypothesis based on the oiled beach sand microbial communities are the subject of **chapter VII**. Next, the characterization of a freshwater meta-community in the Southeast USA monitored for six years in seven locations was utilized to quantify the different biogeographic factors contributing to community assembly. **Chapter VIII** describes the results of this study, documenting distinct microbial provinces within interconnected habitats along the Chattahoochee

River, revealing similarly high impact of seasonality and geographic distance on community variation and extant diversity, and documenting a modest effect of landscape and only a minor effect of other environmental factors in community assembly.

CHAPTER 1: BYPASSING CULTIVATION TO IDENTIFY BACTERIAL SPECIES

Originally published on March 2014 in *Microbe* 9 (3): 111-118[1].

Luis M. Rodriguez-R & Konstantinos T. Konstantinidis.

Summary

Culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species.

- The current approach to defining bacterial species, based on genetic and phenotypic distinctiveness, is problematic.
- Bypassing cultivation to assess natural populations provides a valuable and perhaps more authoritative approach to identifying and defining bacterial species.
- Natural microbial communities are predominantly composed of sequence-discrete populations, with exceptions likely to be found within habitats that undergo frequent fluctuations or for organisms with unique ecologic characteristics.
- Sequence-discrete populations could be given candidate species names until appropriate isolates with ecologically relevant phenotypic properties are characterized.

---

[1] In-text references are discouraged by the original publisher. Instead, a list of references is provided as suggested reading.

- The mechanisms maintaining species, and perhaps more importantly, the relative importance of the mechanisms for different organisms and habitats, are not understood and demand further study.

Whether bacterial species exist as a natural unit remains an unresolved issue, one with important practical challenges, including that of correctly identifying microorganisms and diagnosing the causative agents of microbial diseases. The current bacterial species definition is based on genetic and phenotypic distinctiveness of organisms grouped under the same name.

However, the standard methods used to identify bacterial species, including 16S rRNA gene sequence analysis, DNA-DNA hybridizations, and phenotypic tests under laboratory conditions, can lead to two ecologically and genetically distinct microorganisms being assigned to the same species. For example, *Escherichia coli* isolates can differ by as much as one-third of their genomes and represent important pathogens (*e.g.*, O157:H7 lineage) or nonpathogenic, commensal organisms (*e.g.*, MG1655 lineage).

Another important limit to our ability to distinguish bacterial species is that current approaches test bacterial isolates in the laboratory for phenotypic properties that may differ greatly from natural conditions. Such testing may not assign those isolates into clusters that are representative of natural populations. Instead, the isolates frequently fall along a continuum, a result that poses major challenges for any system that aims to assign organisms to distinct taxa. However, whether that continuum reflects a natural pattern or, instead, an artifact of the methods used and/or the cultivation biases is difficult, if not impossible, to determine. Therefore, it is possible that organisms with distinct ecologies and preferred habitats and/or genotypes are being grouped together incorrectly.

In contrast, assessing organisms in their habitat (*in situ*) enables one to observe credibly natural diversity patterns. Thus, by assessing natural populations and bypassing cultivation bias, culture-independent genomic approaches or, more simply, metagenomics can provide valuable insights into microbial species.

Microbes Form Sequence-Discrete Populations

From our review of findings from large-scale metagenomic studies during the past five years, we came to realize that microbial communities are predominantly organized in sequence-discrete populations. These populations become evident after comparing the genome sequence of one member of the population against those of all co-occurring organisms in that same sample or habitat. Members within the same population share high sequence identity, ranging between 94 and 100% genome average nucleotide identity (ANI).

This range depends on the age of the population, with younger populations showing lower sequence diversity. Those belonging to a particular population show significantly less genetic identity to other co-occurring populations, typically less than 80-85% ANI (genetic discontinuity; Figure 1-1). Members of such a population also tend to show similar abundances among themselves, based on how many metagenomic reads map on each genome sequence, indicating that they are ecologically homogeneous. In contrast, members of different populations, even closely related ones, typically show different abundances, indicating that they are ecologically differentiated.

Within connected habitats having similar conditions, the same sequence-discrete populations are found. In other words, when populations are being dispersed between similar habitats, those populations are likely to remain indistinguishable. These principles are based in part on our analyses of populations within five freshwater lakes in the Southeastern United States that connect with one another via the Chattahoochee River (Figure 1-2).

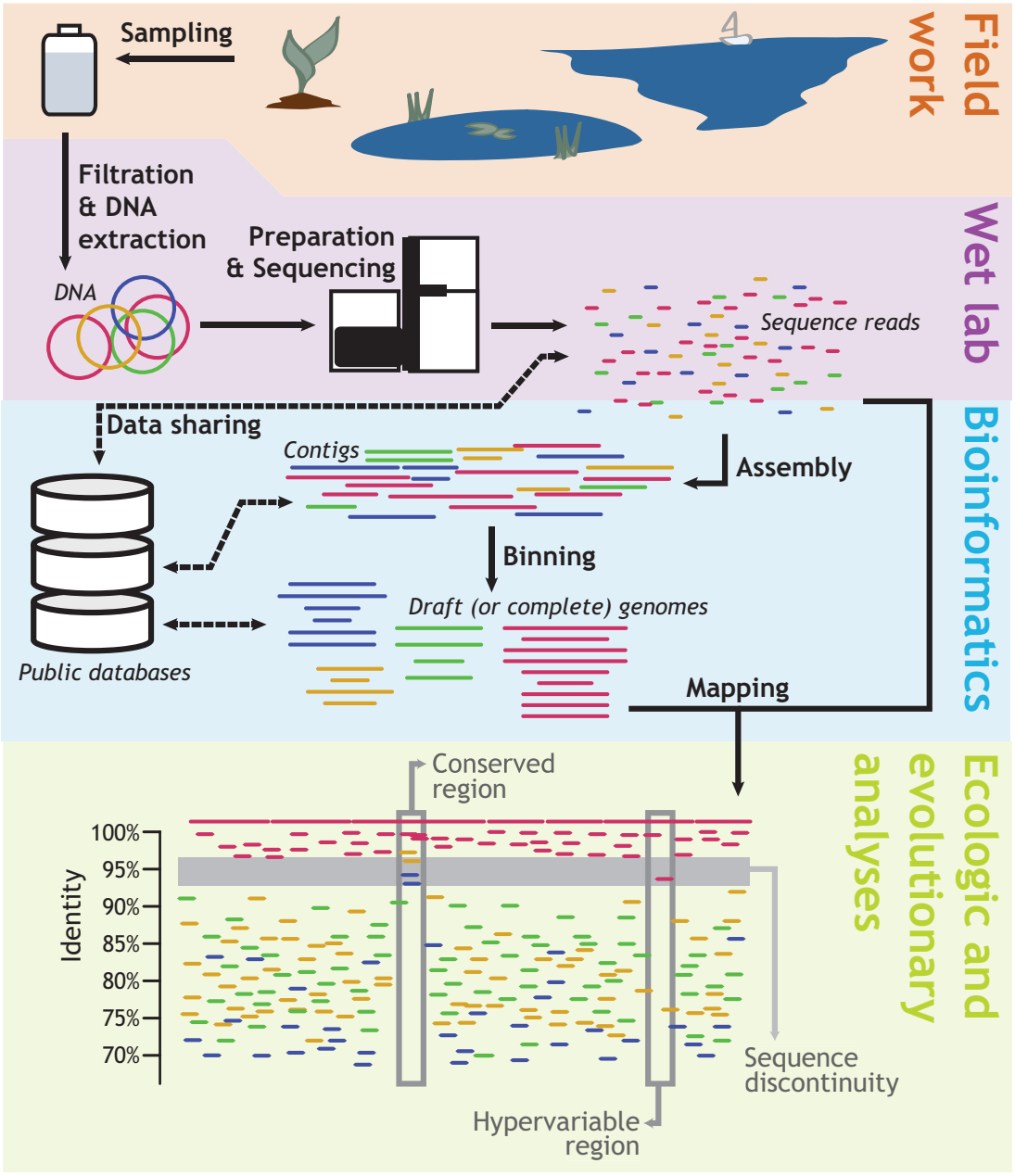


Figure 1-1. Schematic of the metagenomic pipeline to identify sequence-discrete populations.

Reads from metagenomic sequencing of microbial community DNA can be assembled into consensus genomic sequences of cells belonging to the same population. Contigs originating from the same population can be identified based on their sequence characteristics and then grouped into nearly closed draft population genomes (binning). When the original reads of the metagenome are mapped against the contigs of a

reference population (recruitment analysis; bottom), it becomes apparent that each population is sequence-discrete compared to its co-occurring populations. In this hypothetical example, reads originating from members of the reference population (red) evenly match the assembled contigs that represent the population with high nucleotide sequence identities (>97%). In contrast, reads from other populations (other colors) match the reference contigs at lower sequence identities, forming a sequence discontinuity ("gap") in the recruitment plot. Areas that deviate from this pattern are limited to highly conserved regions of the genome (e.g., rRNA operons), where reads from related but distinct populations are recruited due to their highly sequence identity to the reference sequences, or regions characterized by intrapopulation heterogeneity, which typically show lower coverage.

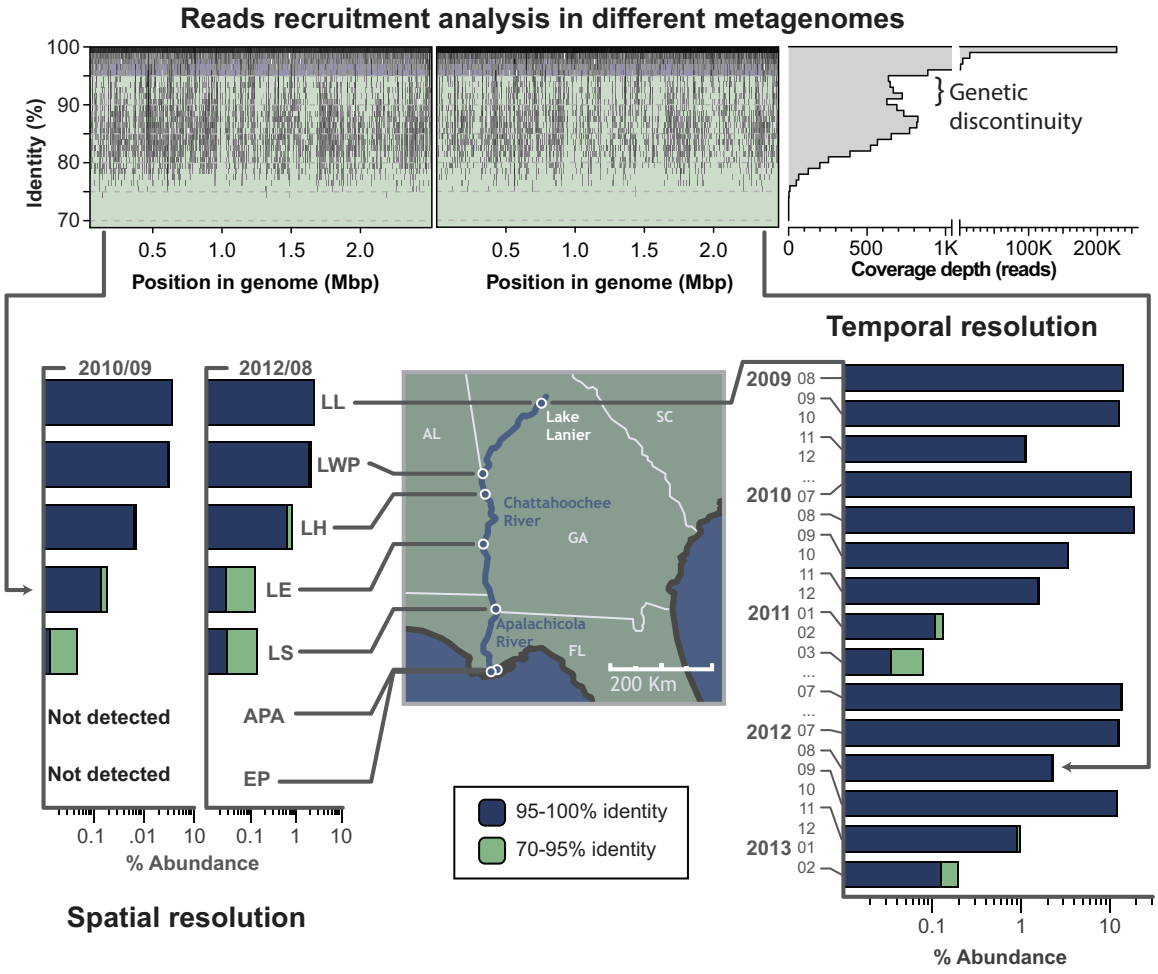


Figure 1-2. Tracking sequence-discrete populations over time and space.

A representative example of a population tracked in the collection of time series metagenomic datasets from lakes in the Southwest United States is shown (Illumina, 100-bp-long reads). The population is an uncultivated member of *Burkholderiacaea*. The abundance of the population and its relatives was quantified in each metagenome by recruiting the reads against the genome sequence of the population, similar to **Figure 1-1**. Reads recruited at >95% nucleotide sequence identity represent the target population (denoted by blue color), while those showing between 70 to 90% identity represent closely related but distinct populations in the same sample (denoted by green color). Note that this population typically shows high abundance in Lake Lanier throughout the year (0.1–10% of the total community), with maxima during the summer months, in four consecutive years, and that its relative population(s) show much lower abundance (about 200 times less). The population is also consistently present in the other lakes along the Chattahoochee River and is no longer identifiable in the estuarine metagenomes, indicating that it is freshwater-adapted. Abbreviations: LL-Lake Lanier; LWP-Lake West Point; LH-Lake Harding; LE-Lake Eufaula; LS-Lake Seminole; APA-Apalachicola Bay; EP-East Point Bay

Our findings suggest that these microbial populations are not ephemeral, clonal amplifications of one or a few cells. Instead, they are long-lived entities that may encompass substantial genetic diversity. Moreover, non-discrete populations are rare, at least for the abundant members of natural communities that can be robustly assessed by metagenomics, and typically ephemeral, as they are associated with regular environmental perturbations such as the mixing of distinct populations that are adapted to living in different depths in the sea caused by ocean upwelling.

More generally, natural microbial communities are predominantly composed of sequence-discrete populations, with exceptions likely to be found within habitats that undergo frequent fluctuations or for organisms with unique ecologic characteristics. Identifying such exceptions will help us to better understand the

ecological and molecular mechanisms that drive the diversity patterns of populations described above.

The mechanisms may involve genetic exchange among members of a bacterial population that keeps them consistent, analogous to sex in higher eukaryotes. On the other hand, ecological coherence, in which different organisms occupy the same niche, coupled with population selective sweeps when a significant genomic innovation takes place, may drive population cohesiveness. Further analysis of sequence-discrete populations will lead to a fuller understanding of how discrete bacterial populations are maintained, interact, and evolve within communities. In any case, such sequence-discrete populations are important units within natural microbial communities.

Comparing Metagenome and Conventional Species Definitions

The sequence-discrete populations identified by metagenomics partly overlap with those encompassed by the conventional species definition. For instance, most –but not all– named bacterial species encompass organisms that show 95% or higher ANI among themselves (Figure 1-3). This level of relatedness contains the greatest diversity for the oldest populations recovered in metagenomes. Hence, the standard definition used for defining species encompasses the sequence-discrete populations. However, the latter tend to show lower genetic intrapopulation diversity than do conventionally named bacterial species.

Analysis of sequence-discrete populations helps us to better recognize several limitations in the conventional approach to defining bacterial species. Members of a population tend to show smaller gene-content differences among themselves – typically, less than 5% of their total genes– compared to named species such as *E. coli*. This trend is consistent with the idea that the former represent more

ecologically and genetically uniform clusters than those encompassed by the current species definition.

In contrast, several named species include organisms isolated from, and hence adapted to living in, different habitats or hosts. For instance, the depth-stratified photosynthetic *Procholoroccus marinus* (*Cyanobacteria*) or the ammonia-oxidizing Marine Group I *Thaumarchaeota* sp. (*Archaea*) likely perform the same metabolism at every depth they inhabit, based on the gene content recovered in the corresponding metagenomes. However, their populations are sequence-discrete, and hence not interchangeable, at different depths due to genomic adaptations to light intensities and hydrostatic pressures that they encounter at their preferred depths.

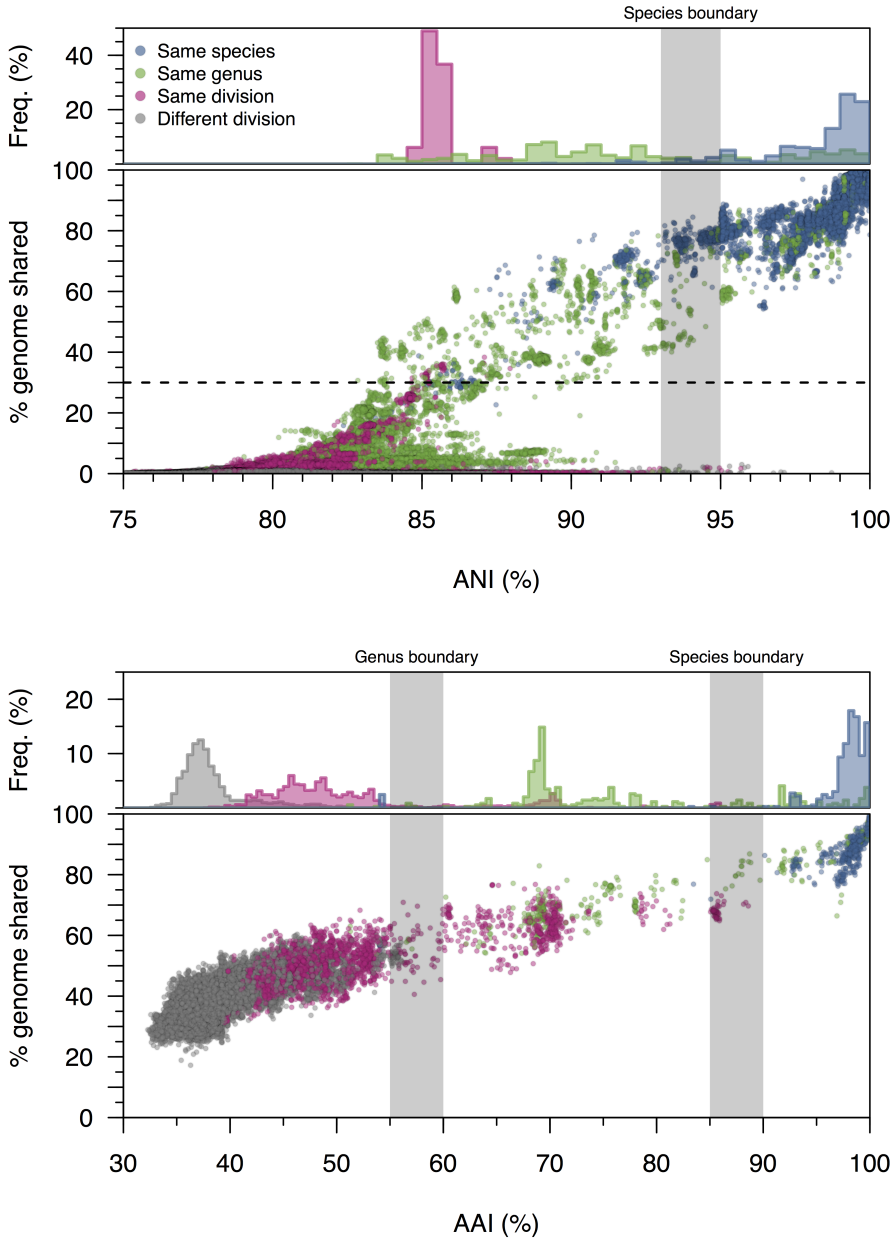


Figure 1-3. Interrelationship between shared gene content and ANI or AAI for bacterial genomes.

ANI/AAI values of all available completed bacterial genomes were computed in pair-wise mode (x axes) and are plotted against the % of genes in the genome shared between the two genomes in the pair (y axes). The analysis shows that ANI offers robust resolution between genomes that share 80 –100% ANI, i.e., within species or among closely related species, and that species that share less than 80% ANI and/or 30% of

their gene content are too divergent to be compared based on the ANI measurement. For the latter genomes, AAI provides a much more robust resolution and should be used instead. Note that a few genomes that share less than 30% of their gene content show higher than 80% ANI due to a few highly conserved genes in the genome or recent horizontal gene exchange, but not because they are highly related evolutionarily.

Similarly, several pathogens encode the same pathogenicity factors and cause similar symptoms in humans or animals. However, these factors are encoded within different genomic backgrounds, as is the case for several lineages of *E. coli*.

From a taxonomic perspective, these subpopulations of the marine species, or the subpopulations of the animal pathogens, should be assignable to the same species because they are characterized by the same phenotypic or metabolic properties that matter to us. From a bacterial perspective, however, they are not interchangeable and each occupies different ecological niches. These findings argue for adopting a more ecological way of defining bacterial species than our current system allows, while also suggesting that modern culture-independent analytical techniques may provide a better way of describing species.

Moving Forward: How To Taxonomically Describe the Uncultivated Majority

Describing a new species depends in part on having a diagnostic phenotype based on traditional biochemical and physiological laboratory techniques as well as showing adequate genetic distinctiveness. Such diagnostic phenotypes typically are not available for sequence-discrete populations of uncultivated organisms.

The taxonomic status *Candidatus* provides a way around this hurdle. Sequence-discrete populations could be given candidate species names until appropriate isolates with ecologically relevant phenotypic properties are characterized. Describing candidate species should be relatively easy for most of the sequence-

discrete populations because they can be identified and tracked based on sequence data, which then can provide means for developing probes with which to analyze cell morphology and other characteristics.

Moreover, single-cell genomic techniques could complement, perhaps even substitute for, shotgun metagenomic efforts because they can recover the complete, or almost complete, genome sequence of microorganisms under study. The *Candidatus* species is perhaps the only pragmatic approach for describing bacterial diversity in nature. If applied systematically to uncultivated microorganisms, it also would greatly facilitate communication among scientists.

Metagenomics data can reveal important gene-content differences or genomic adaptations among sequence-discrete populations. These differences could account for important phenotypic differences *in situ*. For instance, differential usage of amino acids in proteins to cope with hydrostatic pressure differences account, at least in part, for the ability of distinct Marine Group I *Thaumarchaeota* populations to occupy different depths in the oceans. Such differences are impossible to reproduce in the laboratory based on traditional methods or phenotypic assays.

Similarly, environmentally adapted *E. coli* strains possess ecologically important genecontent differences compared to their enteric counterparts. Yet, these organisms are indistinguishable based on traditional phenotypic tests, apparently because those tests do not target ecologically appropriate genes and pathways. In such cases, the differences revealed by genomics and metagenomics can be taken as adequate “phenotypic” differences for delineating species and guiding the design of discriminative phenotypic tests.

As culture-independent transcriptomics and proteomics techniques are further refined, they may better help in assessing the activity of natural microbial populations and, thus, in defining species- or population-diagnostic signatures. In

our experience, transcriptomics and proteomics data typically corroborate population-specific signatures revealed by metagenomics data.

Conclusions, Recommendations, Challenges

Natural microbial communities are predominantly composed of sequence-discrete populations that possess attributes expected for species, which contrasts with a genetic continuum observed between several named species. The discrepancy is presumably attributable to biases introduced by cultivation and human-centered ways of analyzing diversity. Therefore, Bacteria and Archaea appear to form discrete biological units, similar to eukaryotes, and these units are partly encompassed by the current definition of species. Omics data can help to further refine the species definition.

Metagenomic fragment recruitment and ANI provide a reliable means for assessing sequencediscrete populations and determining the level of intrapopulation genetic diversity. For recruitment plots, it is important to use a genome sequence or assembled contig from the same population or a highly related population. ANI can also discriminate between closely related populations (sharing at least 70-75% ANI), offers higher resolution than 16S rRNA gene or multilocus sequence analysis, and is less error-prone and more portable than is the DNA-DNA hybridization method. For more distantly related populations, the average amino acid identity (AAI) should be used because resolution is progressively lost at the nucleotide level.

New species descriptions should be accompanied by their ANI and/or AAI relatedness values, and population relative abundance and persistence over time in situ, assessed by metagenomics or other culture-independent technique. The *Candidatus* species description provides a reliable means to identify and characterize populations with no sequenced representatives, and the ANI values of such populations can be reliably computed based on assembled genome

sequences from metagenomics or single-cell techniques. To facilitate such efforts, we have developed online implementations of the ANI and fragment recruitment tools (available through http://enve-omics.gatech.edu/).

Meanwhile, the mechanisms maintaining sequence-discrete populations (and species), and perhaps more importantly, the relative importance of those mechanisms for different organisms and habitats, are not understood and demand further study. To this end, characterizing isolates of several sequence-discrete populations could help to guide further use of omics data and how to better interpret traditional phenotypic measurements.

Suggested reading

(Fraser, Hanage, & Spratt, 2007; Gevers et al., 2005, 2005; Goris et al., 2007; Konstantinidis, Braff, Karl, & DeLong, 2009; Konstantinidis & Stackebrandt, 2013; Luo et al., 2011; Luo & Konstantinidis, 2011; Oh et al., 2011; Shapiro et al., 2012)

Acknowledgements

K.T.K. is indebted to Jim Tiedje for useful discussions related to the species issue and Ed DeLong for exposing him to the science of metagenomics. Our work is supported in part by the U.S. DOE Office of Science, Biological and Environmental Research Division (BER), Genomic Science Program, Awards No. DE-SC0006662 and DE-SC0004601, and by the U. S. National Science Foundation under Award No 1241046.

CHAPTER 2: A USER'S GUIDE TO QUANTITATIVE AND COMPARATIVE ANALYSIS OF METAGENOMIC DATASETS

Originally published on September 2013 as chapter 23 of *Microbial Metagenomics, Metatranscriptomics, and Metaproteomics* (ed. Edward F. DeLong), in *Methods in Enzymology* 531: 525-547, DOI: 10.1016/B978-0-12-407863-5.00023-X.

Chengwei Luo[1], Luis M. Rodriguez-R[1] & Konstantinos T. Konstantinidis.

Metagenomics has revolutionized microbiological studies during the past decade and provided new insights into the diversity, dynamics, and metabolic potential of natural microbial communities. However, metagenomics still represents a field in development, and standardized tools and approaches to handle and compare metagenomes have not been established yet. An important reason accounting for the latter is the continuous changes in the type of sequencing data available, for example, long versus short sequencing reads. Here, we provide a guide to bioinformatic pipelines developed to accomplish the following tasks, focusing primarily on those developed by our team: (i) assemble a metagenomic dataset; (ii) determine the level of sequence coverage obtained and the amount of sequencing required to obtain complete coverage; (iii) identify the taxonomic affiliation of a metagenomic read or assembled contig; and (iv) determine differentially abundant genes, pathways, and species between different datasets. Most of these pipelines do not depend on the type of sequences available or can

[1] These authors contributed equally to the work.

be easily adjusted to fit different types of sequences, and are freely available (for instance, through our lab Web site: *http://www.enve-omics.gatech.edu/*). The limitations of current approaches, as well as the computational aspects that can be further improved, will also be briefly discussed. The work presented here provides practical guidelines on how to perform metagenomic analysis of microbial communities characterized by varied levels of diversity and establishes approaches to handle the resulting data, independent of the sequencing platform employed.

Introduction

Culture-independent whole-genome shotgun (WGS) DNA sequencing has revolutionized the study of the diversity and ecology of microbial communities during the last decade (Handelsman et al., 2007). However, the tools to analyze metagenomic data are clearly lagging behind developments in sequencing technologies, and several important bioinformatic challenges remain (Hugenholtz & Tyson, 2008; Kunin, Copeland, Lapidus, Mavromatis, & Hugenholtz, 2008). For instance, metagenomic studies of environmental samples typically recover only short (e.g., < 10 kb long) fragments of the genome, which only rarely contain rRNA genes, the backbone of bacterial identification and taxonomic classification (Boone, Castenholz, & Garrity, 2001), either because of chance (< 0.1% of the genome is represented by rRNA genes) or the high similarity among rRNA genes from distinct organisms that prevents their correct assembly from metagenomic data (C. S. Miller, Baker, Thomas, Singer, & Banfield, 2011). Accordingly, identifying and studying novel taxa based on metagenomic approaches remain challenging due to the lack of appropriate non-rRNA-based methods and reference genomes.

Most metagenomic surveys to date have sampled only a small fraction of the total diversity within the target community, especially in highly complex soil/sediment microbial communities (Delmont, Simonet, & Vogel, 2012; Tyson et

al., 2004), and the amount of additional sequencing required to cover the whole diversity has typically remained speculative. Generally speaking, this fraction is termed coverage and depends on both the sequencing effort and the diversity of the microbial community in the sample. Incomplete coverage does not prevent researchers from reaching valuable conclusions about the communities under study, but it constitutes a source of uncertainty and limits several downstream analyses such as assessing the importance of low-abundance (rare) community members. Estimating the diversity of a sample in terms of the number of species or operational taxonomic units (OTUs) present is challenging. Current approaches mainly rely on the construction of rarefaction curves (or similar approaches) based on the identification of OTUs (e.g., (Caporaso et al., 2010; Schloss & Handelsman, 2005)). The application of these techniques in short-read datasets, however, requires either the use of a reference database (to assign/recruit reads to reference sequences and then cluster reference sequences in OTUs) or clustering of assembled sequences (reference-free approach). The former is biased by the limited number of reference genes available in the databases, with the probable exception of the 16S rRNA gene (Cole et al., 2010). The latter is biased by the use of phylogenetic markers that are much more conserved than the average gene in the genome (in order to be sufficiently similar to allow clustering/alignments) such as the ribosomal rRNA genes. However, important levels of genomic and ecological differentiation frequently underlie identical 16S rRNA gene sequences (Acinas et al., 2004; Konstantinidis et al., 2006).

A related challenge for metagenomics is how to identify differentially present genes, pathways, and species between datasets. The issue is complicated not only by the low coverage achieved in typical metagenomic datasets but also by the difficulty in defining microbial species (hence, OTUs are typically preferred instead; reviewed in (Caro-Quintero & Konstantinidis, 2012, p.; Gevers et al., 2005; Rosselló-Mora & Amann, 2001)) and the short-read length of current next-

generation sequencing (NGS) technologies, which limits identification and quantification of target phylogenetic markers. For instance, short-read (i.e., 50–200 bp) NGS technologies have become increasingly popular due to their high throughput and low cost per sequenced base, but it remains unclear whether these technologies can be used to routinely and robustly assemble complete gene and/or individual genome sequences from complex communities. The low coverage typically achieved in metagenomic studies also represents a major challenge for assembly, in addition to short-read length (Delmont et al., 2012). NGS technologies are also changing continuously and, thus, their sequencing errors and artifacts need to be examined, and the associated bioinformatic pipelines to be updated, on a regular basis (e.g., (Luo, Tsementzi, Kyrpides, Read, & Konstantinidis, 2012)).

Here, we describe the bioinformatic approaches others and we have developed to achieve several of the tasks mentioned earlier, focusing primarily on “how to” accomplish the tasks and the limits of each approach, depending on the coverage obtained, type of sequencing technology employed, and objective of the study. A fundamental concept underlying our own approaches is the sequence-discrete populations. Our recent review and synthesis of the major metagenomic studies performed to date on various populations and habitats revealed that natural microbial communities are predominantly composed of discrete populations, with the intrapopulation sequence diversity typically ranging between ~95% and ~100% genome-aggregate average nucleotide identity, or gANI, depending on the population considered (Caro-Quintero & Konstantinidis, 2012, p.; Konstantinidis & DeLong, 2008). The 95% gANI level corresponds tightly to 70% DNA–DNA hybridization, which is commonly used to demarcate bacterial species (Goris et al., 2007). Whether or not these populations should be equated to species remains unclear (Caro-Quintero & Konstantinidis, 2012, p.), but the 95% gANI level appears to represent robust means to define populations, and hence, OTUs. Accordingly, we employed 95% gANI as needed during our

analyses, and our pipelines employ genomic relatedness measures such as gANI, which offers important advantages compared to traditional approaches based on rRNA genes for the same purposes.

How to Assemble a Metagenomic Dataset

Due to the large difference between the desired sequence length for analysis (e.g., the average bacterial gene length is 950 bp and a typical *Escherichia coli* genome is around 4.5 Mbp) and the sequencing read length provided by NGS (e.g., frequently < 200 bp long), assembling WGS reads is usually the first step of metagenomic studies and represents the foundation for various downstream analyses. It is a critical yet challenging step, largely due to short read length and the fact that metagenomes represent mixtures of different genotypes, some closely related to each other. The objective is to obtain assemblies with long average contig length (typically measured by N50, which is defined as the longest length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs) and high quality (e.g., low frequencies of chimeras and base call errors).

Depending on the number of taxa present in the target community lacking sequenced representatives, the assembly process may be reference guided or *de novo*, or a mixture of both. When appropriate representatives are available (we recommend > 90% gANI between reference genome and target population), reference-guided assembly is usually optimal. For instance, more than 3000 reference genomes are currently, or will be soon, available as part of the Human Microbiome Project (HMP; www.hmpdacc.org). Therefore, to reconstruct bacterial genotypes from human microbiome datasets, it is common to first prepare a nonredundant set of reference genome sequences at a given clustering threshold (e.g., 95% gANI) and then map the metagenomic reads to these references using mapping tools such as BLAST (Altschul et al., 1997), BLAT (W. J. Kent, 2002), MAQ (H. Li, Ruan, & Durbin, 2008), or Burrows–

Wheeler transformation-based algorithms, which are suitable for fast short-read mapping (e.g., Illumina or ABI SOLiD platforms), including BWA (H. Li & Durbin, 2009) and Bowtie (Langmead & Salzberg, 2012). Reads mapped to a reference can then be binned together for population assembly (see below), substantially reducing complexity and hence, improving assembly quality. A second round of assembly can be subsequently applied, as necessary, in which the resulting contigs are assembled together with all reads in an attempt to recover genomic islands present in the target population but absent from the reference genome (and thus, missed during the reference-guided assembly).

A more challenging scenario occurs when few or no available references exist and thus,*de novo* assembly is needed, as is often the case for metagenomes from most natural habitats. In general, *de novo* assemblers fall into two categories: overlap-based and graph algorithm-based (J. R. Miller, Koren, & Sutton, 2010). The former perform well with long sequences such as those generated by Roche 454 and Sanger sequencers (Luo, Tsementzi, Kyrpides, & Konstantinidis, 2012). Exemplary assemblers of this category include the widely used Newbler package (Margulies et al., 2005), Celera assembler (Myers et al., 2000), and Arachne (Batzoglou et al., 2002). Graph algorithm-based assemblers have recently gained popularity for metagenomic studies mainly due to the prevalence of short-read sequencing data. Early generation assemblers from this category employed greedy algorithms and include SSAKE (Warren, Sutton, Jones, & Holt, 2007), VCAKE (Jeck et al., 2007), and SHARCGS (Dohm, Lottaz, Borodina, & Himmelbauer, 2007). They were later outperformed by *de Bruijn* graph-based algorithms such as Velvet (Zerbino & Birney, 2008), Euler (Chaisson, Brinza, & Pevzner, 2009; Chaisson & Pevzner, 2008), SOAPdenovo (R. Li et al., 2010), ABySS (Simpson et al., 2009), and AllPaths (Butler et al., 2008), each of which builds its core data structure using different variations of a *K*-mer graph. A *K*-mer graph is composed of nodes, which are short nucleotide sequences (*K*-mers), and edges, which connect the nodes. Transitional

relationship is a commonly implemented approach to connect two *K*-mers; for example, the 4-mer ATGA can transition to the 4-mer TGAC by removing the 5′-end letter A and adding the 3′-end letter C.

The latter tools were originally designed for assembly of single genomes, not metagenomes; and several properties of metagenomes violate basic assumptions of the corresponding algorithms. For instance, Velvet assumes even coverage along the target genome—an assumption that does not hold true for metagenomes, where the relative abundance of each species is almost always different (uneven). Several methods were more recently developed to overcome these limitations. For example, Meta-IDBA (Peng, Leung, Yiu, & Chin, 2011) and MetaVelvet (Namiki, Hachiya, Tanaka, & Sakakibara, 2012) tackle the problem by first isolating graphs into components that likely belong to the same population (or coverage bin) and then performing a variant-tolerating assembly for each individual component. In our previous study, we developed a robust hybrid protocol that combines the power of *de Bruijn* graph algorithms (Velvet and SOAPdenovo) and overlap-based approaches (Newbler package) to provide higher-quality assemblies, with larger N50 values (Luo, Tsementzi, Kyrpides, & Konstantinidis, 2012). In short, this protocol first removes redundancy among preassembled contigs from several independent runs of Velvet for preferably its metagenomic variant MetaVelvet (Namiki et al., 2012), and SOAPdenovo using a wide range of *K*-mers from 21 to 63 (three runs per algorithm are recommended) and then combines and assembles the remaining contigs into final contigs using Newbler. This hybrid protocol showed a twofold increase in average contig length and returned about 50% more assembled reads, while maintaining similar assembled sequence quality when compared with assemblies solely constructed using Velvet or SOAPdenovo in various metagenomes, including freshwater planktonic (Oh et al., 2011) and ocean beach sand samples (Rodriguez-R et al., 2015).

The ultimate goal in metagenome assembly is to recover whole-genome sequences. Initially, these efforts were focused on relatively simple communities, such as the acid mine drainage system (Denef & Banfield, 2012; Simmons et al., 2008), archaeal symbionts of marine sponges (Hallam et al., 2006), and the gut microbiome of premature infants (Morowitz et al., 2011), and combined manual inspection with popular assembly software. It is important to note that in all these studies, a mosaic genome representing the average genome of the target population, rather than a single genotypic variant present in the sample, was recovered.

More recently, successful assembly of genomes from complex communities using more automated approaches has been reported. For instance, Iverson and colleagues were able to recover nearly completed genomes of marine *Euryarchaeota*, *Thaumarchaeota*, and *Flavobacteria*, each representing 4–10% of a surface water metagenome, by using paired-end read information of a jumping library (insert size 2–3 kb as opposed to the typical ~ 300 bp size) to link precontigs (Iverson et al., 2012). Wrighton and colleagues recovered 49 incomplete genomes from groundwater metagenomes (completeness level varied between 41% and 95%) spanning different phyla by integrating self-organizing, map-based sequence binning methods and iterative coassembling techniques (Wrighton et al., 2012). As metagenomic sequencing becomes more and more affordable, related metagenomes along spatial and temporal (i.e., time-series) gradients will be increasingly common. Hence, there is a need to develop robust genome assembly methods suitable for time-series metagenomes.

In our experience, it is nearly impossible to obtain assemblies that resolve the genomes of closely related, co-occurring genotypes (strains) of the same population for almost any sample or method employed as the intrapopulation divergence is usually too small for assemblers to differentiate, frequently at the same level as sequencing errors. However, in some special cases of low-diversity communities or deep-branching populations (i.e., no close relatives co-

occurring in the community), it has been possible to resolve a small set of target gene sequences at the strain level with the aid of visualization tools such as Strainer (Eppley, Tyson, Getz, & Banfield, 2007) or computationally intensive (C. S. Miller et al., 2011) expectation–maximization (EM) algorithm-based methods such as EMIRGE (C. S. Miller et al., 2011). We have also recently performed a comprehensive *in silico* evaluation of the strain level resolution of our hybrid assembly protocol for different scenarios of intrapopulation genetic structure, and the reader is directed to the original publication for further details (Luo, Tsementzi, Kyrpides, & Konstantinidis, 2012).

Metagenomic assemblies should be carefully evaluated before being used for further analysis. Factors that can affect assembly quality include the intrinsic characteristics of the target community (e.g., richness and evenness of species, G + C% content and size of genomes, and abundance of repeated sequences), the experimental design (e.g., sequencing throughput, library size, and the choice of sequencing platform), and the parameter settings of the assembler. To evaluate the effect of these factors, simulated systems have been extensively used. For instance, Charuvaka and Rangwala evaluated the relationship between assembly quality (e.g., chimera frequency, average contig length) and community complexity and choice of *K*-mer size using simulated reads (Charuvaka & Rangwala, 2011). In our previous study, we spiked-in real Illumina reads of isolate genomes into Illumina metagenomes, instead of using simulated reads (Luo, Tsementzi, Kyrpides, & Konstantinidis, 2012), and this approach provides reliable means to evaluate several parameters of the assembly (Figure 2-1). The evaluation showed that, with about 20 × coverage of the target population, its genome can be recovered at high-draft status (Branscomb & Predki, 2002), while at lower coverage levels, chimeric contigs increase in frequency and probably cause, in part, community diversity to be (artificially) overestimated. The relationships among population coverage and different types of sequencing errors and artifacts were examined more thoroughly for 100-bp-long Illumina

data, and the reader is referred to the original publications for further details (Luo, Tsementzi, Kyrpides, & Konstantinidis, 2012; Luo, Tsementzi, Kyrpides, Read, et al., 2012).

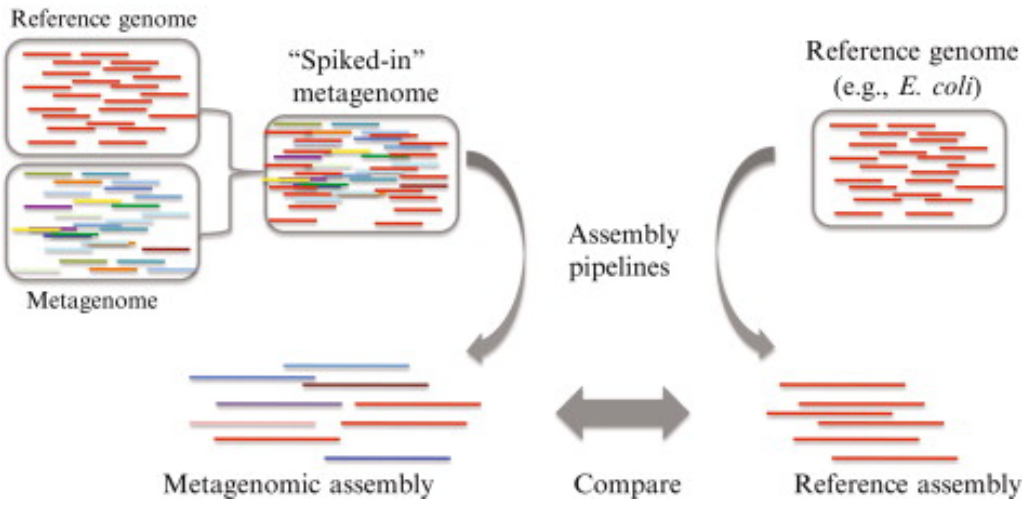


Figure 2-1. An approach to assess assembly parameters and output based on *in silico*-generated "spiked-in" metagenomes.

To assess the impact of assembly parameters (e.g., *K*-mer, consensus cutoff, and minimum coverage), reads of reference genome(s), generated ideally from the same sequencing platform as the target metagenome, are spiked into the metagenome to form an *in silico* dataset. This *in silico* dataset is sequentially assembled, and the assembled contigs are compared against the reference genome sequence or the contigs assembled from the reference genome data alone. Note that a similar approach can be used for other purposes, for instance, to assess tools for functions other than assembly.

How to Determine the Fraction of the Community Captured in a Metagenome

Almost all metagenomic datasets published today lack completeness, that is, they do not capture all the DNA molecules or species within the target community. Several approaches have been devised to assess the level of

community coverage, depending on the specific aim of the study and the additional information available about the community.

Single species analysis

In the simplest case, the objective is to determine the whole-genome coverage of one or a few target species or, more precisely, the coverage breadth, to differentiate from the average number of times each position is sequenced (called sequencing depth and sometimes referred to as coverage depth); the captured fraction of the remaining genomes of the community is not relevant. In 1988, (Lander & Waterman, 1988) provided a model aiming to solve this problem. In this model, the coverage breadth of a genome can be predicted from the sequencing depth. The latter can be easily estimated as the total size of the dataset multiplied by the abundance of the target species and divided by its genome size. For example, if a species constitutes about 4% of the community, with a genome of 4 Mbp, the expected sequencing depth in a metagenome of 300 Mbp would be 3X. That is, each position of the genome is sequenced three times, on average. Next, the following expression can be applied to predict the coverage breadth:

$$Cov_{breadth} = 1 - e^{Cov_{depth}}$$

Equation 2-1

For the given example above, the expected coverage breadth will be 95%. Although the Lander–Waterman model is still widely used, more refined models have been proposed (e.g., (Wendl, 2006; Wendl et al., 2001; Wendl, Kota, Weinstock, & Mitreva, 2012), but no software implementations of the latter models are currently available.

Whole community analysis based on single gene markers

A more complex, but also more common, scenario is to target a specific gene marker, as opposed to a given taxon, to assess its total diversity and/or new variants in the sample. Typical examples include the analysis of rRNA genes, which can also provide taxonomic information, and thus, estimations of the total number of taxa in the community. All gene-based analyses have a common step, that is, to cluster sequences into OTUs. This allows the application of a well-known family of indexes derived from the nonparametric Good's coverage estimator (Good, 1953). Good's estimator has been shown to be statistically efficient (Esty, 1986), and yet it benefits from its simplicity:

$$\tilde{C} = 1 - \frac{n_1}{N}$$

Equation 2-2

where n_1 is the number of clusters (or OTUs) found with only one observation and N is the total number of clusters or taxa, if clustering was based on rRNA gene sequences, found. Alternatively, one can estimate the total number of clusters or taxa (as opposed to coverage) using the Chao1 index (Chao, 1984), which is derived from Good's estimator. This is remarkable, because it means that the number of clusters in the unobserved part of the community can be estimated from the distribution of the observed clusters. Most practical applications, however, only marginally benefit from capturing extremely rare variants. In other words, the objective is usually to capture near-completeness as opposed to full completeness. This difference may seem subtle but, because of the long-tail distribution of clusters within many communities, it can translate to several orders of magnitude difference in sequencing effort. A more useful representation for this purpose are collector's curves (also referred to as rarefaction curves), which are derived from the capture–recapture methods of

population ecology and essentially represent plots of the number of clusters observed as a function of sample size. Mothur (Schloss et al., 2009) is a popular package to generate collector's curves from sequencing data, but the same or similar methods have been implemented in a wide variety of packages and libraries. Essentially, these curves allow visual inspection of the level of sample saturation by sequencing because a more pronounced plateau (if any) indicates more saturated sampling (hence, near-completeness). This concept was recently generalized as "complexity curves" in the package preseq (Daley & Smith, 2013), providing a method to accurately project these curves to completeness, assuming sampling is unsaturated and the curve is close enough to saturation for the projection to be reliable.

Whole community analyses based on whole genomes

A more challenging scenario is to estimate the coverage at the whole-genome level in shotgun metagenomes, where the genome size and abundance of each member of the community are typically inaccessible and clustering is not feasible. There are three main solutions to this problem. The first is to approximate the distribution of abundances and genome sizes. This technique is exemplified by the work of (Stanhope, 2010), and for the most part by the work of (Hooper et al., 2010), and represents an intuitive solution but current implementations largely depend on the availability of optimal assemblies, making it inappropriate for very small datasets and/or very complex communities. A second solution is to identify gene markers sufficiently general to allow characterization of the captured portion, and subsequently apply techniques derived from targeting single gene markers (discussed earlier). For example, clusters can be formed using rRNA gene sequences extracted from the shotgun datasets. Alternatively, genomic-based taxonomic classifications can be used to define clusters. However, the latter approach depends on a comprehensive database of reference genomes, which is typically not available for most natural communities, uses only a small fraction of the datasets, and is subject to large influence of randomness.

Nonetheless, in both cases described earlier, coverage and richness can be predicted and collector's curves can be constructed and projected.

Finally, intrinsic characteristics of the metagenomes can be used to estimate the captured fraction, sidestepping the biases introduced by suboptimal assemblies or incomplete reference databases. We recently presented Nonpareil (Rodriguez-R & Konstantinidis, 2014c), a method that examines the redundancy among all reads of a metagenome to estimate the average coverage of the genomes in the community. In addition, Nonpareil allows projecting the average coverage at increased sequencing efforts to predict the coverage that could be attained at any given size of dataset. This method enables fast calculation of coverage in entire metagenomic datasets, even those that are several Gbp (giga base pairs) in size, provides estimations of the amount of sequencing required to cover the complete or nearly complete diversity of the sample, and reflects the relative diversity of samples when compared with reference datasets or between samples. Our analyses of both *in silico*-constructed as well as real datasets from the HMP suggest that Nonpareil outperforms other tools for the same purposes and is applicable to microbial communities that show a wide range of diversity and complexity. Nonpareil is available for online querying through http://enve-omics.gatech.edu/ and as a stand-alone binary at https://www.github.com/lmrodriguezr/nonpareil.

How to Identify the Taxonomic Identity of a Metagenomic Sequence

Identifying the taxonomic affiliation of a sequence recovered in a metagenome remains challenging, primarily for the following reasons. First, the current collection of genome sequences is far from comprehensive and, thus, does not represent well the organisms in most natural environments (Cole et al., 2010). Second, no universally accepted definition of bacterial species exists, and hence, it is difficult to decide the degree of novelty of a new taxon (Konstantinidis & Tiedje, 2007). Finally, horizontal gene transfer, which is pronounced in the

microbial world (Gogarten & Townsend, 2005), creates inconsistencies between sequence and organismal phylogeny, further complicating the issue. To tackle this issue, various algorithms have been developed, which can be classified into two categories: composition based and alignment based (Mande, Mohammed, & Ghosh, 2012). The former utilize sequence statistics with robust taxonomic signal; the latter are based on homology searches between query sequences and a database and employ sequence similarity as a proxy for taxonomic relatedness. Each approach has its own advantages and disadvantages, and the choice of which tool to implement often depends on the specific objective of the study.

Composition-based methods

The most important advantage of composition-based methods is that they are almost reference independent (most of them still need reference genomes to train the underlying algorithms) and, therefore, can assign taxonomy to sequences that do not match any reference sequences (unlike alignment-based methods). Also, they are, in general, faster and require less computational resources. Popular composition-based algorithms include, but are not limited to, PhyloPythia (McHardy, Martín, Tsirigos, Hugenholtz, & Rigoutsos, 2007; Patil et al., 2011), NBC (Rosen, Reichenberger, & Rosenfeld, 2011), Phymm (Brady & Salzberg, 2009), RAIPhy (Nalbantoglu, Way, Hinrichs, & Sayood, 2011), and TACOA (Diaz, Krause, Goesmann, Niehaus, & Nattkemper, 2009). However, these methods do not usually perform well on short sequences (e.g., < 500 bp long), largely due to the insufficient information provided by such sequences.

Alignment-based methods

These methods classify query sequences according to their relatedness to available reference sequences, based on the corresponding alignments. The primary alignment engines are BLAST (Altschul, Gish, Miller, Myers, & Lipman,

1990); BLAST-like tools such as BLAT (W. J. Kent, 2002); hidden Markov model-based tools such as HMMer (Finn, Clements, & Eddy, 2011); or Burrows–Wheeler transform-based methods such as MAQ (H. Li et al., 2008), BWA (H. Li & Durbin, 2009), and Bowtie (Langmead & Salzberg, 2012). Alignment-based methods are computationally more expensive. However, they are probably indispensible for every metagenomic study as their results are used for additional downstream analyses such as gene annotation and community profiling. Further, they would assign an increasingly larger number of query sequences as the available reference genome sequences from isolation or single-cell efforts (Stepanauskas, 2012) increase. Alignment-based classifiers include MG-RAST (Meyer et al., 2008), MEGAN (Huson, Auch, Qi, & Schuster, 2007), MARTA (Horton, Bodenhausen, & Bergelson, 2010), CARMA (Krause et al., 2008), AMPHORA (M. Wu & Eisen, 2008), TreePhyler (Schreiber, Gumrich, Daniel, & Meinicke, 2010), and others. We have recently presented MyTaxa[2], an algorithm that employs unique design elements to classify at least 5% more sequences than any existing alignment-based tool ((Luo, Rodriguez-R, & Konstantinidis, 2014); and also available through http://enve-omics.gatech.edu/). MyTaxa is briefly described in the following section.

The MyTaxa algorithm

MeTaxa differs from other alignment-based methods in that it takes into account all genes encoded on a query sequence, weighting each gene based on its (predetermined) classifying power. The weights reflect: (i) how well the gene resolves the classification at a given taxonomic level based on its degree of sequence conservation (e.g., 16S rRNA resolves poorly the species level in

[2] The original publication of this chapter included the name “MeTaxa”, a name we modified to “MyTaxa” after the first publication of this chapter.

contrast to the genus level or higher) and (ii) how consistent the gene phylogeny is with species phylogeny, the latter being approximated by the genome-aggregate average amino acid identity (gAAI). Parameterized weights and alignment-based matches against a reference database are subsequently integrated via a maximum likelihood algorithm. MyTaxa reports the probability for each possible taxonomic classification of the query sequence as well as the degree of novelty for sequences representing novel taxa (e.g., novel species, genus, or phylum) based on previously determined gAAI standards that correspond well to taxonomic ranks (Konstantinidis & Tiedje, 2005b). The standardized approach to assess novelty represents another important improvement provided by MyTaxa compared to previous approaches. The gene weights are precalculated “offline” based on the publicly available completed and draft genomes and are included in the MyTaxa package. Users need only to provide, as input to MeTaxa, a BLAST tabular-like output from the search of each query sequence against their preferred reference database, for example, NR, KEGG, SwissProt, etc. MyTaxa can return high-precision predictions for thousands of input query sequences in a matter of a few minutes on a personal laptop computer.

Combination and optimization

As the two categories of methods have their own advantages and disadvantages, hybrid protocols have been more recently reported. For instance, PhymmBL combines BLAST output (alignment-based) and Phymm algorithm (composition-based) to achieve higher performance (Brady & Salzberg, 2009). In general, alignment-based methods are usually more accurate, while for query sequences without significant matches to the reference database, composition-based methods are probably the only options available. Among the composition-based methods, we have obtained good results with NBC, although the best method of choice would depend on the specific objective of the study and the type of data available.

How to Determine Differentially Abundant Genes, Pathways, and Species

Any standardized annotation of metagenomic sequences, whether it involves genes, pathways, species, or any other functional or taxonomic categorization, can essentially be the subject of comparison across samples. To detect annotations that are differentially abundant between datasets with confidence, a statistical approach is necessary to account for under sampling of community diversity and the stochastic nature of WGS metagenomes. This task, generally referred to as profile comparison, can be divided into three main steps. First, metagenomic sequences must be annotated. Although annotation is feasible for short metagenomic reads, a more reliable annotation is often achieved based on assembled contigs. For example, entire contigs can be assigned a taxonomic affiliation, and predicted genes encoded on the contigs a putative function. Next, the abundance of annotations (features) is determined by mapping the original reads onto the features, generating a table of read counts. Finally, the statistical significance of the differences is evaluated. Several tools have been specifically designed to carry out statistical tests for metagenomic datasets such as the Statistical Analysis of Metagenomic Profiles, or STAMPS, package (Parks & Beiko, 2010). STAMPS can analyze any set of features across sets of metagenomes, in any of three modes: comparison of two samples, comparison of several samples in two groups (e.g., treatment vs. control), and comparison of multiple samples. It should also be noted that there is rich literature on the comparison of differentially abundant features from other types of studies, for example, transcriptomics (RNA-seq and CHiP-Seq), mostly varying on the assumptions about the underlying distribution of counts (e.g., binomial, Poisson, overdispersed Poisson, negative binomial). A guide and an evaluation of several methods were recently presented (Fang, Martin, & Wang, 2012; Schreiber et al., 2010). The type of data and most assumptions used in these fields are essentially the same as in metagenomics; hence, the methods are applicable to the problem discussed earlier.

We have developed a simple and robust statistical method to identify differentially abundant genes, pathways, or organisms between well-replicated control versus treatment metagenomes. In brief, the method combines resampling techniques (Jackknife), the DESeq package (Anders & Huber, 2010), and binomial hypothesis testing. Suppose we have m treatment and n control samples, a Jackknife method is used to generate all possible combinations of $\lfloor m/2 \rfloor$ treatment and $\lfloor n/2 \rfloor$ control samples ($\lfloor x \rfloor$ denotes the floor function of a real number, x, which maps it to the largest previous integer). For each combination, a normalized count table (see below for normalization) is generated by mapping sequences (e.g., reads) to different features (e.g., genes, pathways, population genomes); each row in the table represents a feature and each column represents a sample. DESeq is then applied to detect the difference between treatment and control samples for each feature. For a specific feature (row in the table), the log2 fold changes determined by the DESeq analysis of all combinations of samples follow a distribution; the mean represents the best estimate of fold change and the variance reflects the uncertainty of the estimate. A binomial test is then carried out to test the significance of the log2 fold change (1 for significantly different log2 fold change; 0 otherwise), and the P-value is adjusted for false discovery rate using the Benjamini–Hochberg method (Benjamini & Hochberg, 1995).

Modifications for Other Scenarios

Our method was originally developed to compare well-replicated (6 replicates per treatment at minimum; 10 replicates or more recommended) complex soil metagenomes generated by the Illumina HiSeq platform. To extend it to other types of sequence data or samples characterized by different complexities, modifications might be required, most often at the normalization and sparse counts (i.e., features with zero counts in certain samples) steps, and some limitations could emerge. Count table normalization is necessary in order to compare samples with different sequencing depths; the most popular approach is

to present the number of sequences as a fraction of the total sequences of the corresponding sample. However, the latter approach undermines the statistical power derived from count data and thus is not recommended. Instead, we suggest normalizing counts using quantile-based methods, like the ones described previously (Bullard, Purdom, Hansen, & Dudoit, 2010; Fang et al., 2012). The DESeq algorithm in our above-mentioned approach normalizes samples based on similar methods. Additional normalization steps may be required in some cases, however. For instance, when samples differ substantially in sequencing quality (e.g., different percentage of reads passing quality trimming) or when the sample datasets are too large to determine the number of reads for every feature, a resampling technique should be used to subsample the datasets at random to the same size. For features with low relative abundance, the sparse counts among samples will frequently lead to inaccurate testing results. A pragmatic way to account for this is to set a cutoff on the number of sequences mapped to a feature, and the features with lower counts are discarded from further analysis. To determine an appropriate cutoff, a Fisher's exact test-based method is proposed to simulate the impact of different cutoffs on the accuracy (White, Nagarajan, & Pop, 2009), while an alternative is to modify the hypothesis testing as discussed in (Tusher, Tibshirani, & Chu, 2001). The smaller the number of metagenomes compared (e.g., $n = 2$), the more important is to account for the sparse count issue.

Limitations and Perspectives for the Future

We presented here a practical guide to the analysis and interpretation of metagenomic data that should be useful in future studies across different habitats and microbial groups. It is important to realize, however, that the field of metagenomics is currently undergoing a major expansion, and new tools and approaches are being developed, including pipelines that integrate various tools to offer a comprehensive analysis of metagenomic datasets such as the Kbase project (http://kbase.science.energy.gov/) and MetAMOS (Treangen et al., 2013).

It was not possible to mention all recent developments as part of the present document nor was that our intention. Our goal was instead to provide practical recommendations based on current knowledge and types of sequence data available, and a reference point for future developments (Figure 2-2). We anticipate that several of the approaches and tools described earlier will require modification in the not-so-distant future, mostly due to new types of sequence data that will become available. For instance, it is foreseeable that metagenome assembly will become less challenging when single molecule sequencing, which can provide long sequence fragments on the order of tens of kilobases (Eid et al., 2009; Stoddart, Heron, Mikhailova, Maglia, & Bayley, 2009), becomes more routine. Related to the latter, single-cell technologies, which can provide the draft genome sequence of individual cells in a sample, are becoming increasingly more throughput, reliable (e.g., no DNA contamination), and affordable (Stepanauskas, 2012), and can greatly assist metagenomic studies by providing reference points in the analysis. Integrated approaches that combine multiple omics techniques, including transcriptomics and proteomics, have clear advantages, depending also on the specific objective(s) of the study. For instance, combining shotgun metagenomics with single-cell genomics is advantageous for population genetic studies but probably does not offer as much when the goal is to compare the gene content of different communities or the same community after a perturbation.

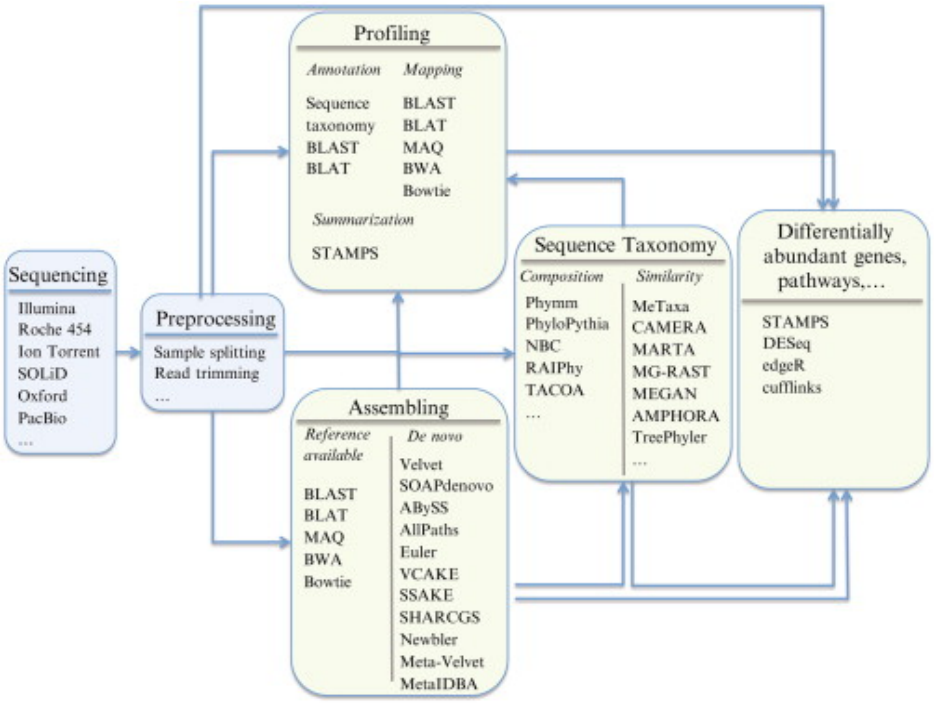


Figure 2-2. A graphical representation of the major components and associated bioinformatic tools of a typical metagenomics study.

The bioinformatic analysis of metagenomic datasets starts with read quality processing (e.g., read trimming, multiplexing barcode removal), followed by community profiling, assembling, sequence taxonomic assignment (individual read or assembled contig level), and finally identification of differentially abundant genes, pathways, and organisms. The graph shows the relevant tools that can be used at each step and the arrays connecting different analysis represent the possible workflow. See text for more details.

Even with the availability of longer sequences or single-cell genomes, however, a major computational challenge remains in how to handle and analyze the increasing volume of data that is produced by such approaches. Clearly, the tools and algorithms available do not scale up with the amount of data produced by new sequencers and single-cell technologies. Innovating solutions in terms of hardware implementations, algorithm optimizations (e.g., (Pell et al., 2012)), and (redundant) data reduction are highly needed to cope with the data available.

Until then, obtaining the complete picture of complex microbial communities that are composed of hundreds to thousands of distinct species will remain somewhat a utopia.

Acknowledgments

We thank Mike Weigand for helpful suggestions regarding the chapter. Our work is supported in part by the U.S. DOE Office of Science, Biological and Environmental Research Division (BER), Genomic Science Program, Award No. DE-SC0006662 and DE-SC0004601, and by U.S. National Science Foundation under Award No. 1241046.

CHAPTER 3: THE ENVEOMICS COLLECTION: A TOOLBOX FOR SPECIALIZED ANALYSES OF MICROBIAL GENOMES AND METAGENOMES

Released on March 2016 in PeerJ Preprints 4: e1900v1,
DOI: 10.7287/peerj.preprints.1900v1.

Luis M. Rodriguez-R & Konstantinos T. Konstantinidis.

Genomic and metagenomic analyses are increasingly becoming commonplace in several areas of biological research, but recurrent specialized analyses are frequently reported as in-house scripts rarely available after publication. We describe the enveomics collection, a growing set of actively maintained scripts for several recurrent and specialized tasks in microbial genomics and metagenomics, and present a graphical user interface and several case studies. Our resource includes previously described as well as new algorithms such as Transformed-space Resampling In Biased Sets (TRIBS), a novel method to evaluate phylogenetic under- or over-dispersion in reference sets with strong phylogenetic bias. The enveomics collection is freely available under the terms of the Artistic License 2.0 at *https://github.com/lmrodriguezr/enveomics* and for online analysis at *http://enve-omics.ce.gatech.edu*.

Introduction

Microbial genomics and metagenomics have become key components of several areas of modern research including biomedicine, epidemiology, plant and animal pathology, environmental engineering and science, microbial ecology, and

evolutionary biology. Specialized computational analyses in these areas are hence becoming commonplace for the non-expert, often resulting in the reimplementation of scripts critical for understanding and reproducing the reported results, with varying levels of quality, reproducibility, and availability. While the literature analysing *ad hoc* scripts is scarce, a 2004 survey on the availability of URLs reported in MEDLINE found that 19% of 1,020 analysed pages were always unavailable, and only 63% were always available (Wren, 2004). Moreover, we searched the manuscripts available as full-text in PubMed Central with the terms “in-house script”, “in-house developed script”, “in-house perl script” (other languages didn’t return additional results), or the same terms in plural, and found 1,929 matching articles (as of January 05, 2016). From these, 1,654 were related to genomics or metagenomics and 449 to microbial genomics or metagenomics. We further explored the latter set of manuscripts, and found that only 6% provided access to the source code (26/449), with an additional 1% reporting websites no longer available (3/449) or not including the reported scripts (3/449). 3% of the manuscripts explicitly indicated that the scripts were available upon request (13/449), but in only 3 cases the authors provided the code within two months of the request. The large majority (90%) did not provide any reference, provided references to previous publications of the same group not including the source code, or referenced only the programming language in which the scripts were implemented. While thissurvey is not a systematic analysis of availability of in-house scripts, nor does it provide quality assessments, the results do underscore the prevalence of a phenomenon that undermines reproducibility in studies applying microbial genomics or metagenomics. On one hand, individual tools are the basis for complete and reproducible methods that are reported either in manuscripts, white papers, or standard operating procedures (SOPs), but the abovementioned statistics showed that the tools infrequently become available. On the other hand, a suitable alternative to providing developed tools for results reproducibility is to make data on each step of the analyses publicly available, but this approach is

also rarely adopted, with the added issues of larger file sizes and heterogeneity, making the documentation of data even more challenging than documenting code. Here, we present a growing collection of actively maintained scripts for several recurrent and specialized tasks in microbial genomics and metagenomics, together with comprehensive documentation, a graphical user interface, and some cases of use. Our collection may also constitute a reference example for other researchers in the future, and an actively maintained framework that could be collaboratively expanded.

Implementation

The enveomics collection is a multi-language set of over 70 independent scripts that accomplish specialized tasks in genomics and metagenomics, including code in Ruby, Perl, AWK, Bash, and R. In addition, the collection features reusable libraries that automate recurrent sub-tasks. For example, the `enveomics_rb` Ruby library includes object-oriented representations of trees, read-placement results, sets of orthologous genes, and complex (non-contiguous) sequence coordinates, together with methods for accessing and downloading remote data. The R code is packaged into a single library (`enveomics.R`) to simplify its distribution.

Preferred file formats

Format incompatibility between data sources and analysis tools is a common problem in bioinformatics, and there are several tools and libraries dedicated to the translation between format specifications (Rice, Longden, & Bleasby, 2000). In order to mitigate the impact of this problem, the enveomics collection has been designed to support only a reduced number of formats, with a wide range of alternative variations. For example, sequence files are expected to be in FastA, but the scripts in the collection always support multi-FastA and can parse variations of the definition lines and colon-lead comments (Suppl. Table S1).

Supported formats for other data types include tabular BLAST for similarity searches (including variations with additional columns and comments), JPlace for phylogenetic read placement (Matsen, Hoffman, Gallagher, & Stamatakis, 2012), and tables in raw text with tab-delimited columns.

Access to remote servers

Local data sources are often insufficient and commonly outdated. In response, we have implemented several utilities to simplify the automated access to remote databases using the Representational State Transfer Application Program Interfaces (RESTful APIs) of the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI), the U.S. National Center for Biotechnology Information (NCBI) E-Utilities, the Kyoto Encyclopedia of Genes and Genomes (KEGG), and the M5nr (Kanehisa & Goto, 2000; W. Li et al., 2015; Sayers et al., 2009; Wilke et al., 2012). All the scripts using these modules are categorized in Annotation/database mapping, and include additional documentation such as informing the user that third-party software or database is used and thus, the latter resources should be cited appropriately in any resulting publications.

Enveomics-GUI

The documentation and parameter descriptions for all the scripts are standardized into a set of JSON files that allow the dynamic creation of Graphical User Interface (GUI) forms though the enveomics-GUI package, including a set of examples and reference files (Figure 3-1). The package is a collection of Ruby libraries, including EnveGUI that implements graphical user interaction with Shoes 4 (https://github.com/shoes/shoes4). The JSON files meet the definitions of the ECMA-404 standard (Ecma International, 2013), but their processing (implemented in the EnveJSON library) ignores object entries with "_" key, that are utilized for comments, and implements external file inclusion using the object

entries with "`_include`" key. The package is distributed as source code (requires Shoes 4 and JRuby), as a stand-alone OS-independent Java Archive (JAR), and as a bundled Mac OS X application.

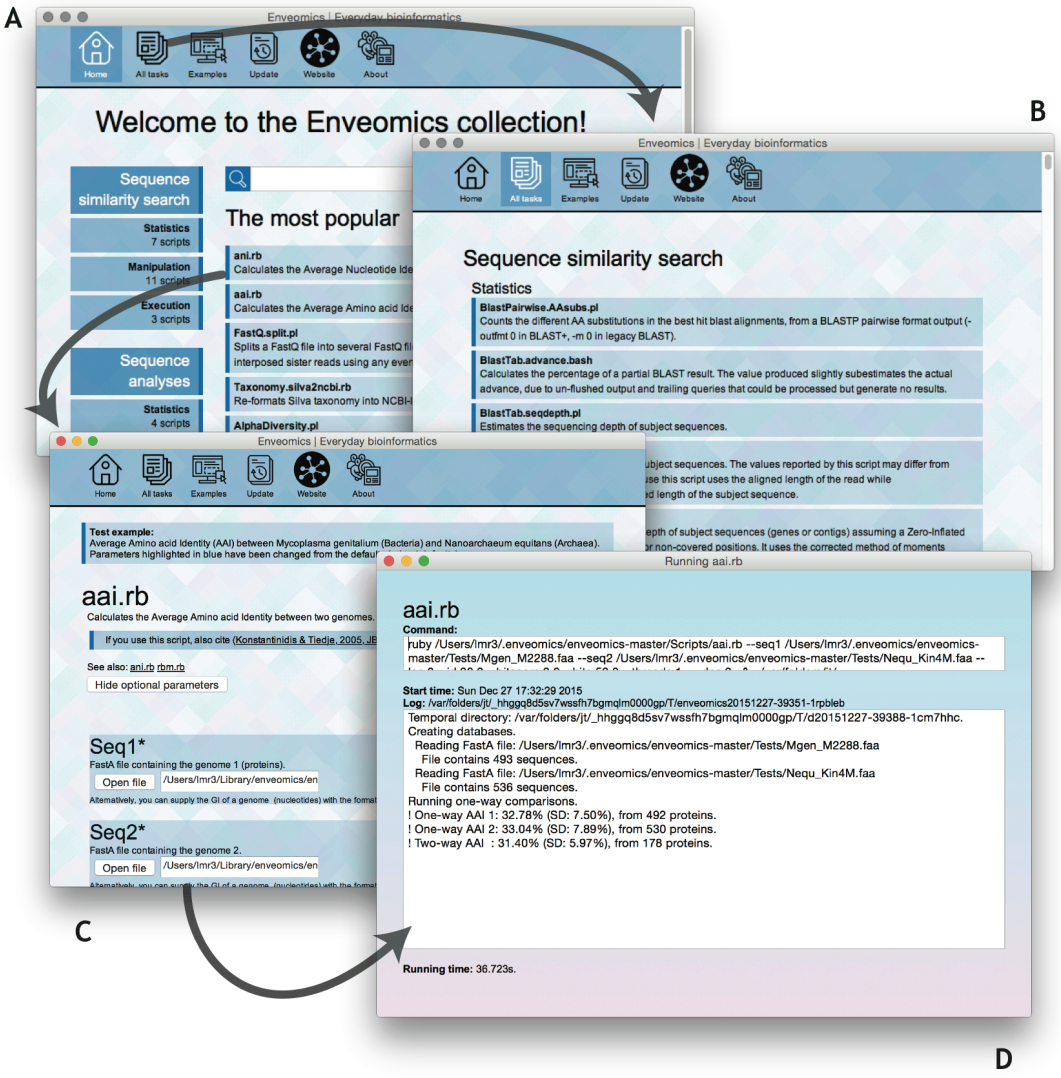


Figure 3-1. Screen captures of the enveomics GUI in Mac OS X.

A. Initial (home) screen with search bar, listing all categories and subcategories, and highlighting a few selected and randomly picked scripts. **B.** Complete list of scripts per category. **C.** Task form for `aai.rb` pre-filled with an example. **D.** Result of the `aai.rb` analysis. All screen captures correspond to v0.1.2. Future versions may differ.

Results

Reimplementations and novel algorithms

The enveomics collection aims to simplify the use of novel and previously described algorithms for the analysis of community (*e.g.*, `Chao1.pl`, `AlphaDiversity.pl`, `Newick.autoprune.R`) and population diversity (*e.g.*, `BlastTab.recplot2.R`, `RecPlot2.find_peaks.R`, `CharTable.classify.rb`), among other tasks in microbial genomics and metagenomics. Here we describe representative modules (see also Suppl. Table S1) including algorithms developed by our group.

Reciprocal Best Match and Average Sequence Identity. The detection of Reciprocal Best Matches (RBMs) is a reliable method for the identification of orthology (Wolf & Koonin, 2012) that has been widely used in genome-aggregate metrics of genetic relatedness (Goris et al., 2007; Konstantinidis & Tiedje, 2005a). Although phylogenetic reconstruction remains the gold standard for orthology detection, RBM provides a fast alternative for high-throughput analyses such as genome-wide scanning. The enveomics collection contains utilities for the detection of RBMs (`rbm.rb`) and the compilation of Orthology Groups (OGs; `ogs.mcl.rb`), as well as the estimation of Average Nucleotide Identity (ANI; `ani.rb`; generally suitable for comparisons of genomes assigned to the same genus) and Average Amino acid Identity (AAI; `aai.rb`; suitable for comparisons of genomes assigned to different species).

Transformed-space Resampling In Biased Sets (TRIBS). Environmental analyses often rely on pre-existing reference databases as a proxy to the presence of features in query datasets. However, databases seldom represent the source of the query sets uniformly, introducing sampling biases. TRIBS is a novel algorithm that reduces the impact of biased sampling by uniformly resampling reference objects in a transformed space generated by

Multidimensional Scaling (MDS). This enables the testing of differences between a dataset and a given subset for the detection of under- or over-dispersion of traits (`TRIBS.test.R`, `TRIBS.plot-test.R`). The method was originally designed for the detection of phylogenetic under-dispersion of traits in groups of genomes with strong phylogenetic bias (Suppl. Fig. S1; `TRIBS.test.R`).

Automated pruning of phylogenetic trees. The enveomics collection also features a utility to automatically prune trees keeping clade representatives (`Newick.autoprune.R`), a useful tool for the navigation of large trees such as those produced from 16S rRNA gene databases. This script iteratively extracts the cophenetic matrix from a tree and removes terminal nodes with at least one other node closer than a target minimum distance (by default, the first quartile of all the paired distances in the initial tree). In some cases, the complete cophenetic matrix is prohibitively expensive to estimate (in the initial iterations for large trees); in those cases the script takes a random sample of terminal nodes and removes sister nodes (or their children) closer than the target distance. An example of a pruned tree is presented in Figure 3-2 B-C.

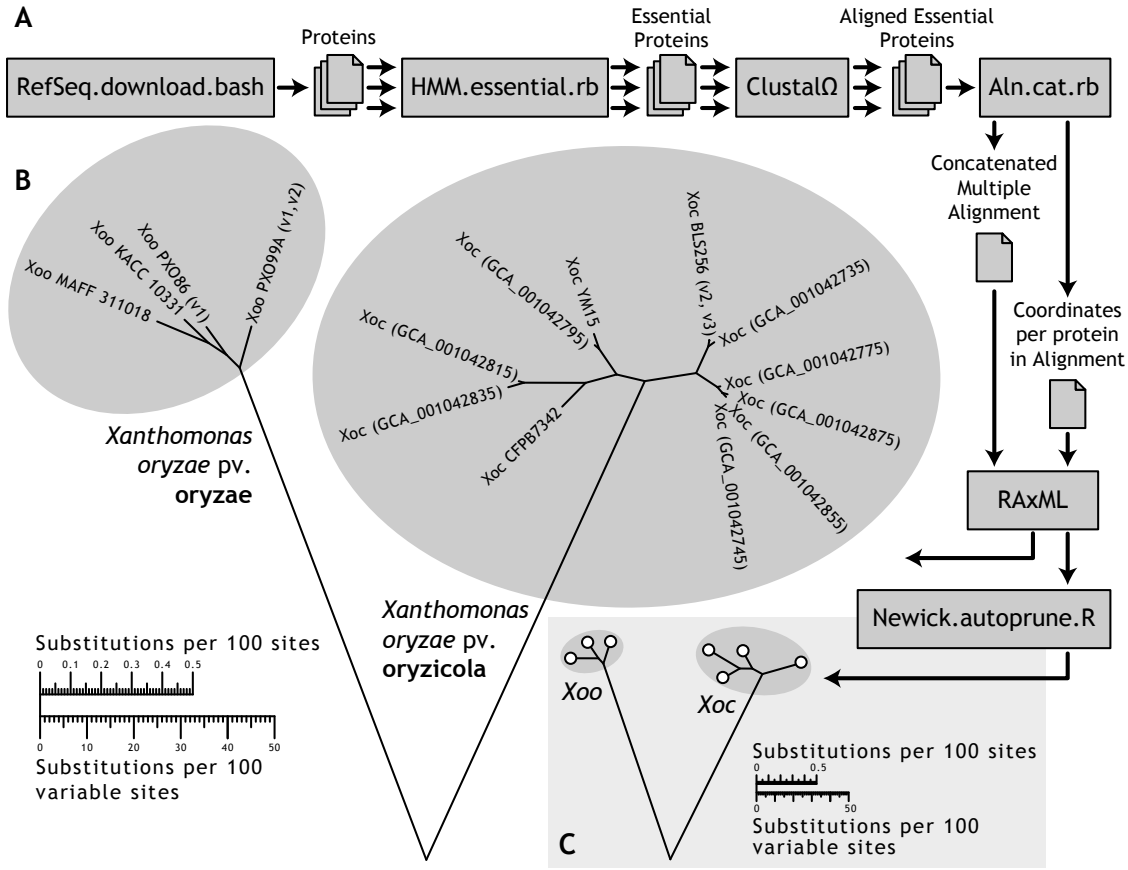


Figure 3-2. Example of a complete workflow primarily using tools from the enveomics collection applied to *Xanthomonas oryzae* genomes.

A. The workflow uses the enveomics collection, Clustal Omega, and RAxML, to generate a phylogenetic tree based on the concatenated alignment of 105 single-copy essential genes. **B.** In the resulting phylogeny, two clades form naturally corresponding to the pathovars oryzae (*Xoo*, left) and oryzicola (*Xoc*, right). Note that the tree is un-rooted, but the rooting point is suggested (vertex) based on phylogenomics of the genus (Rodriguez-R et al., 2012). The invariable sites were removed using `Aln.cat.rb` (35,386 sites removed) and the phylogeny was reconstructed using the remaining 552 informative sites. From the 105 detected essential genes, 22 were identical across all genomes and were excluded from the analysis. **C.** A simplified version of the tree was produced by automatically pruning terminal nodes at a distance lesser than 0.01, resulting in a tree with 7 genomes (out of 17) with similar structure.

Case studies using the enveomics collection

Core genome phylogenies. Whole-genome phylogenetic reconstruction is a powerful method for the resolution of evolutionary relationships. The enveomics collection includes utilities to download genomes of a given species, detect RBMs between pairs of genomes, identify OGs, and identify the genes shared among all the genomes in the collection –the core genome– (`RefSeq.download.bash`, `rbm.rb`, `ogs.mcl.rb`, `ogs.extract.rb`). After computing independent alignments of each core OG, a concatenated alignment can be generated with the options of excluding invariable sites and keeping track of coordinates (`Aln.cat.rb`) to generate robust phylogenies with OG-specific models. In addition, the OGs can be used to estimate several gene-content properties (`ogs.stats.rb`) and the rarefied core and pan-genomes (`ogs.core-pan.rb`) of the species. As a less expensive alternative to the entire core genome phylogeny, one could also identify and analyse only the collection of 111 single-copy genes typically present in archaeal (often ~26 genes) and bacterial (often ~106 genes) genomes (`HMM.essential.rb`). We implemented a workflow using the enveomics collection, together with the multiple alignment tool Clustal Omega (Sievers et al., 2011) and the phylogenetic reconstruction tool RAxML (Stamatakis, 2014) and applied it to the 17 publicly available complete genomes of *Xanthomonas oryzae* (Figure 3-2). The resulting phylogeny identifies known pathovars and the overall structure is consistent with a previous phylogenomic reconstruction (Rodriguez-R et al., 2012). The complete analysis is fully automated, and the code is deposited in the enveomics collection at `Examples/essential-phylogeny.bash`. The execution took 31.2 minutes using two 2.9 GHz processors.

Gene variants in a metagenome. Characterizing the allelic diversity of genes in metagenomes allows targeted analyses of specific traits and the exploration of population discreteness and intra-population variations, independent of

cultivation and amplification (Caro-Quintero & Konstantinidis, 2012; Rodriguez-R & Konstantinidis, 2014a). We explored the intra-population diversity of a metagenomic-recovered bin (LL-70.1) using the mapping of metagenomic reads (LL_1101B; SRR948448 (Tsementzi, Poretsky, Rodriguez-R, Luo, & Konstantinidis, 2014)) from a water sample in January 2011 at Lake Lanier (GA, USA). Read mapping was performed with BLAST (Altschul et al., 1990), and results were analysed and visualized using `BlastTab.catsbj.pl` and `BlastTab.recplot.R` (Figure 3-3), revealing small gene-content variations (panels 2 and 4), but a large allelic variation and the presence of closely related organisms at about 90% ANI (panels 1 and 3). However, a clear genetic discontinuity exists separating this species, as evidenced by the gap around 95% identity, a phenomenon further discussed in (Caro-Quintero & Konstantinidis, 2012; Rodriguez-R & Konstantinidis, 2014a). The enveomics collection also includes utilities for the normalization (`BlastTab.topHits_sorted.rb`, `BlastTab.sumPerHit.pl`, `BlastTab.seqdepth_ZIP.pl`), characterization (`Chao1.pl`, `AlphaDiversity.pl`, `TRIBS.test.R`), and visualization (`Table.barplot.R`, `TRIBS.plot-test.R`, `BlastTab.recplot2.R`) of reference allele distributions in a metagenome using read mapping. Additionally, the allelic diversity of a particular gene of interest can be explored beyond known variants using phylogenetic read placement (Berger, Krompass, & Stamatakis, 2011; Matsen, Kodner, & Armbrust, 2010), that can be visualized in the interactive Tree of Life (iToL) (Letunic & Bork, 2007), and further explored to characterize distances to known variants or ancestral nodes (`JPlace.to_iToL.rb`, `JPlace.distances.rb`) as in (Rodriguez-R et al., 2015).

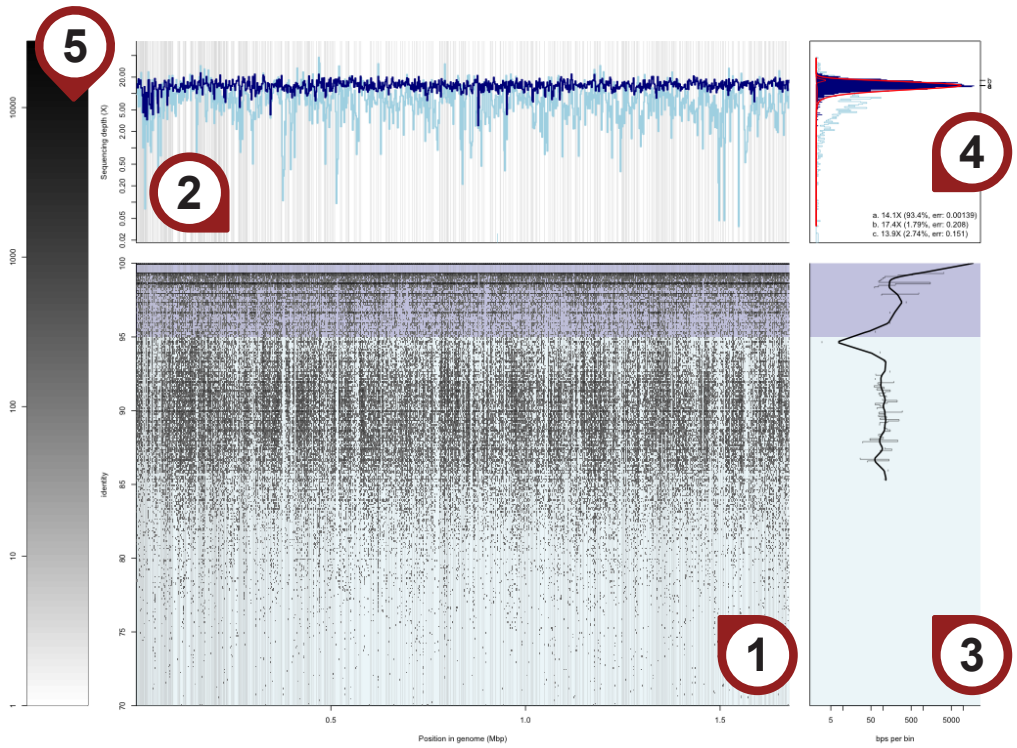


Figure 3-3. Example of a fragment recruitment plot.

This figure showcases the result of processing a BLAST search of metagenomic short sequencing reads (150 bp long in this case; each matching read is represented by a dot in main panel 1) against a population genome sequence assembled/binned from the same metagenome (X-axis). The tabular BLAST result was parsed using `BlastTab.catsbj.pl`, and graphical representation was generated with the `BlastTab.recplot2.R`. The circled numbers 1 through 5 denote the distinct panels of the layout: (1) Main panel representing the reads recruited, placed by location (X-axis) and identity (Y-axis). (2) Sequencing depth across the reference, in logarithmic scale. Bars at the bottom represent regions without mapping reads (sequencing depth of zero). (3) Identity histogram of mapping reads (light gray) and smoothed spline (black), in logarithmic scale. (4) Sequencing depth histogram. Peaks from values above 95% identity are automatically identified as skewed normal distributions (red), with centrality measures, percentage of the reference length, and fit error (bottom-right legend) reported for each peak (marked in the right edge). (5) Color scale for the number of stacked reads per 2-dimensional bin in panel 1. The background of panels 1 and 3, and the line colors in panels 2 and 4, correspond to matches with identity above (dark blue)

and below (light blue) a user-defined cutoff. By default, the identity cutoff is set to 95%, corresponding to the species boundary (Konstantinidis & Tiedje, 2005a). See also (Rodriguez-R & Konstantinidis, 2014a) for additional discussion.

Availability

The source code for all the scripts and additional documentation are deposited and maintained at https://github.com/lmrodriguezr/enveomics. The enveomics-GUI is maintained at https://github.com/lmrodriguezr/enveomics-gui. In addition, we have made available a server with online interfaces for select tools at http://enve-omics.ce.gatech.edu/, including the ANI and AAI calculators, and previously reported tools like Nonpareil (Rodriguez-R & Konstantinidis, 2014c), a tool to estimate the level of coverage in metagenomic samples, and MyTaxa (Luo et al., 2014), a taxonomic classification tool for sequence fragments.

Discussion

The enveomics collection offers a wide array of tools implementing specialized recurrent tasks in microbial genomics and metagenomics and is aimed for users with or without expertise in bioinformatics. The collection features **(i)** a web-based interface for select tools and the complete documentation of all the tools, **(ii)** a comprehensive graphical user interface (GUI), **(iii)** a command-line interface (CLI) that allows integration with development platforms and automation, and **(iv)** Ruby and R application interfaces (API) for developers. In addition, the collection has a language-agnostic design, allowing the implementation of different tools in the most convenient language depending on available libraries or other considerations. To allow this heterogeneity, all the tools are integrated using a standardized JSON-based documentation scheme, allowing the incorporation of additional tools into the collection for the different interfaces. Finally, examples of input data and parameters are provided to encourage the quick use of the tools without dauntingly extensive user manuals.

A few of the scripts in our collection, in particular those implementing the most simple tasks, are overlapping with those developed by others (*e.g.*, (Cock et al., 2009; Rice et al., 2000; Stajich et al., 2002)). Our goal here was not to perform exhaustive comparisons to previously published scripts. As explained above, these scripts were frequently not available for comparisons. Rather, the goal was to put together a resource that offers easy-to-use tools for the non-bioinformatician and is comprehensive with respect to recurrent tasks in microbiome research. As such, we hope that the scientific community will find this resource useful, and will provide feedback on the scripts and algorithms, and suggestions for further improvements.

Supplementary data

Supplementary data are available at https://peerj.com/preprints/1900.

Acknowledgment

We would like to thank Luis H. (Coto) Orellana, Despina Tsementzi, and Michael R. Weigand, whose active involvement in testing and valuable discussions made this collection possible, and Dr. James R. Cole for his commentary on the tools described and the present manuscript. We would also like to thank early users for their active feedback.

Funding

This work was supported by the United States Department of Energy, Biological and Environmental Research Division (BER), Genomic Science Program [DE-SC0006662, DE-SC0004601]; and the United States National Science Foundation [1241046, 1356288].

CHAPTER 4: ESTIMATING COVERAGE IN METAGENOMIC DATA SETS AND WHY IT MATTERS

Originally published on November 2014 in *ISME Journal* 8 (11): 2349-2351,
DOI: 10.1038/ismej.2014.76.

Luis M. Rodriguez-R & Konstantinos T. Konstantinidis.

A 'metagenome' is the theoretical collection of genomes from all members of a given microbial community, and a 'metagenomic data set' is the subset captured in a given sequencing event. Although these terms are often used interchangeably and metagenomic data sets are regularly called metagenomes by synecdoche, their relationship is analogous to sample and population in statistics. The fraction of the metagenome represented in the metagenomic data set, termed coverage (not to be confused with the repetition of features, termed sequencing depth), is of key importance in assessing statistical significance of features sampled (taxa, genes and so on). However, quantitative computational methods to assess the level of coverage are limited, a problem we have recently attempted to solve. In extreme cases, where small data sets are used to characterize complex communities, misleading inferences can arise. For instance, random variation can be frequently mistaken for real differences in comparisons of metagenomic data sets with extreme differences in coverage. Further, insufficient coverage also reduces the detection limits and statistical power of the comparisons, hiding real, ecologically relevant trends and differences (Figure 4-1). We demonstrate here how available solutions can determine the level of sequencing coverage obtained by metagenomic data sets and thus, guide their robust analysis and comparison.

One widely used qualitative method to estimate coverage is a rarefaction curve, sometimes also called a collector or complexity curve. This method relies on the observation that the curve of rarefied counts of any feature (for example, operational taxonomic units, named species, predicted genes, functional categories or even short motifs) should plateau if the sample is close to saturation. Use of rarefaction curves in microbial community studies was popularized by tools such as mothur (Schloss et al., 2009) and recently extended to include accurate projections at higher sequencing efforts by preseq (Daley & Smith, 2013), which allows the estimation of coverage across features (arithmetic mean). However, this technique and others like it typically rely on a high-quality assembly, comprehensive reference data sets or both, which are often unavailable for complex or poorly characterized communities (with the probable exception of ribosomal RNA (rRNA) genes). Moreover, the preseq projection is optimized for single-species data sets and, therefore, does not scale for mixtures of species, making it insufficient for accurate estimations with complex metagenomic data sets. Without accurate projections, rarefaction curves can only be used to determine whether a data set is close to saturation, a useful but insufficient assessment of coverage. Performing this task with rRNA genes is also problematic; largely because their high sequence conservation frequently masks important levels of genetic and ecological differentiation among closely related organisms (Caro-Quintero & Konstantinidis, 2012).

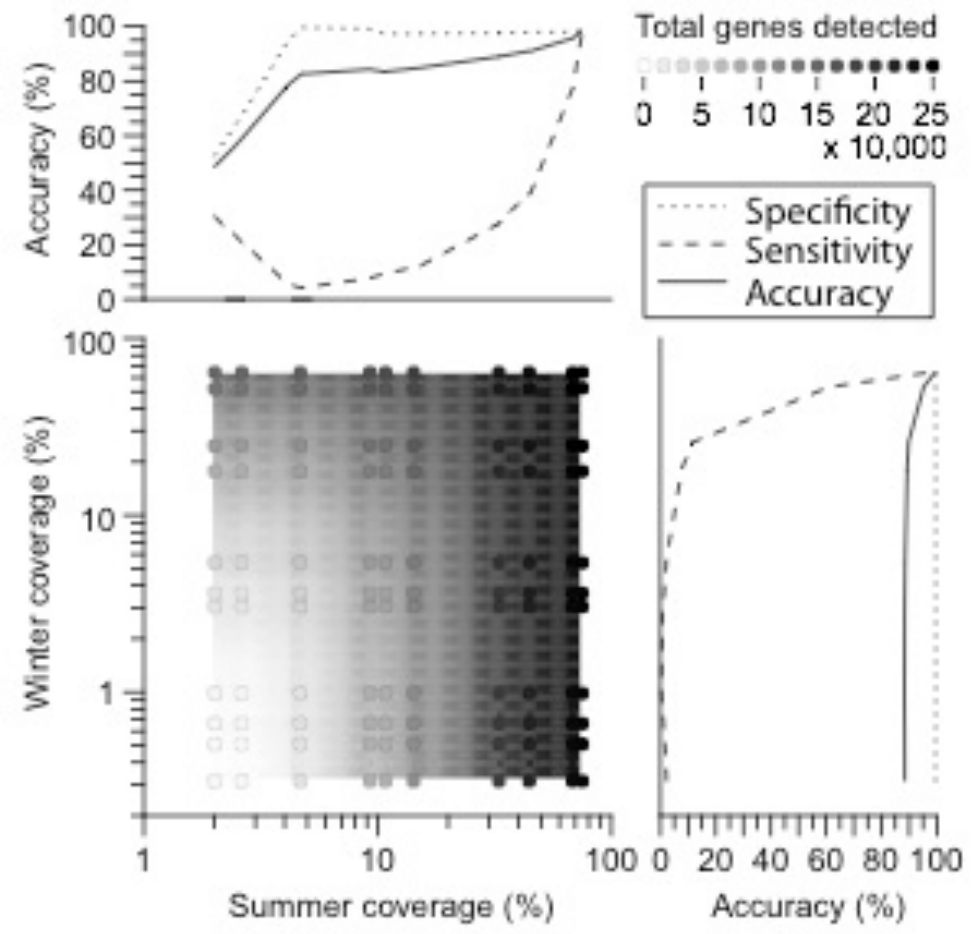


Figure 4-1. Effect of average coverage on detection of differentially abundant features.

The abundance of nonredundant genes (assembled and clustered at 98% amino-acid identity) detected in the metagenomes of Lake Lanier (Atlanta, GA, USA; Sequence Read Archive Projects SRP028408, SRP005437-9; abundance estimates were based on read-mapping at 95% nucleotide identity) was compared between three summer and two winter samples, at different levels of subsampling (0.01–50% of the total data set) and the coverage was computed using Nonpareil (coverage axes). The main panel (bottom-left) shows the number of detected genes, represented by the color of the circles (see legend). The values between subsamples were estimated using bicubic interpolation. Note that the detection of genes is more strongly affected by the coverage in summer data sets owing to lower gene richness in the winter data sets. The additional panels correspond to the comparisons of the subsamples against the complete (not

subsampled) data sets, which showed 64% and 75% coverage for winter and summer, respectively. The comparison between complete winter and summer data sets (top-right circle in main panel) was used as a reference for the definition of true/false positives (TP/FP) and true/false negatives (TN/FN). Sensitivity was defined as TP/(TP+FN), specificity as TN/(FP+TN) and accuracy of the test as (TP+TN)/(TP+TN+FN+FN). Sensitivity, specificity and accuracy were interpolated using cubic splines with smoothing parameter 0.6. Differential abundance was defined as adjusted *P*-value ≤ 0.1 in the negative binomial test implemented in DESeq (Anders & Huber, 2010). Note that sensitivity drops rapidly when coverage of any (or both) of the collections of data sets decreases, while specificity is typically high, except at extreme differences in coverage. In general, the accuracy was compromised (<90%) in data sets with >twofold difference of coverage.

Another approach is to estimate the coverage of one or a few target species in the metagenomic data set using simple statistical approaches such as the Lander–Waterman expressions (Lander & Waterman, 1988), while ignoring the remaining genomes of the community. Such methods are useful in studies targeting specific species in a community in order, for instance, to recover complete genomes. The main drawbacks of this approach include a lack of implemented software and the requirement of reliable estimates for genome size and abundance of the target species, which often poorly represent the community as a whole. No matter how limiting this approach may appear, it can be applied to many available metagenomic data sets, is based on robust statistical frameworks (Wendl et al., 2012) and the interpretation of coverage is straightforward: breadth of the genome covered by sequencing reads.

Finally, genome-wide approaches that capitalize on community modeling and/or modeling of contig sequencing depth have been proposed (for example, (Hooper et al., 2010; Stanhope, 2010)). Such approaches are independent of comprehensive reference databases, which broadens their applicability, but depend on assumed abundance (and genome size) distributions and high-quality

assemblies. Moreover, no software has been available to facilitate their application to real metagenomes.

We recently presented Nonpareil (Rodriguez-R & Konstantinidis, 2014c) as an alternative approach. Nonpareil examines redundancy among the individual reads of a whole-genome shotgun metagenomic data set to quantitatively assess the abundance-weighted average coverage of the data set (for example, Figure 4-1). Therefore, Nonpareil is independent of assembly, reference databases or abundance distribution models, and allows for direct comparisons between data sets and with other quantitative metrics. Furthermore, it projects the average coverage at larger sequencing efforts, providing an estimate of the amount of sequencing required to reach any given coverage and means to quickly rank diversity in metagenomic data sets before assembly or taxonomic classification (Figure 4-2). Finally, Nonpareil uses empirical cutoffs to determine redundant reads, which represent well the area of genetic discontinuity frequently observed among the sequence-discrete populations that typify natural microbial communities based on previous metagenomic surveys (Caro-Quintero & Konstantinidis, 2012). Accordingly, Nonpareil does not distinguish between subpopulations. It is important to note, however, that Nonpareil estimates are based on the organisms recovered in a metagenomic data set, that is, they represent abundance-weighted values, analogous to how metagenomic data sets preferentially represent the abundant organisms in a sample. Thus, in cases where the goal is to characterize all members of the community, or rare members preferentially, and most of these members are not represented in the metagenomic data set due to very high species richness and/or relatively low sequencing effort, Nonpareil estimates may be limited, and should be complemented with genome- or marker-based estimations.

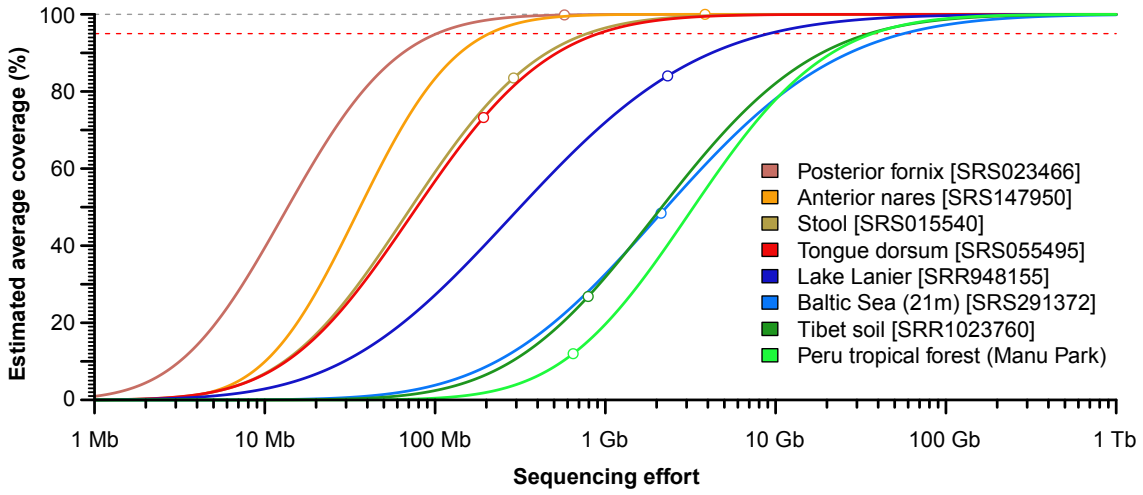


Figure 4-2. Comparison of diversity and coverage in available metagenomic data sets using Nonpareil curves.

The abundance-weighted average coverage is presented as a function of sequencing effort in the form of Nonpareil curves (Rodriguez-R & Konstantinidis, 2014c) for selected available metagenomic data sets. Note that more diverse communities require larger sequencing efforts to achieve the same level of coverage, hence located rightward in the plot. Four samples of the Human Microbiome Project are shown that represent communities in the human microbiome of varying diversity, all of which are less diverse than selected environmental samples. Soil (Tibet soil and Peru tropical forest) and marine (Baltic sea, 21 m depth) samples are the most diverse among those selected. The Sequence Read Archive identifier of each sample is provided within squared brackets, except for the Peru tropical forest sample obtained from (Fierer et al., 2012).

Using Nonpareil, we were able to directly compare the abundance-weighted average coverage of subsampled data sets with frequent analyses in microbial ecology studies. Fewer genes were identified as differentially abundant between data sets with lower coverage at a nearly log-linear rate (Figure 4-1, main panel), and both the significance and power of the statistical test decreased in these cases. The sensitivity of the tests rapidly declined as coverage decreased, while the specificity experienced a dramatic drop when comparing data sets with

extremely different coverage, indicating a high rate of false positives (Figure 1-1, smaller panels). In general, we have observed that data sets with average coverage above 60% perform better in terms of assembly and detection of differentially abundant genes (see also (Rodriguez-R & Konstantinidis, 2014c)), and comparisons of data sets with extreme differences in coverage (for example, >twofold) should be avoided.

Here, we advocated for the estimation of the average coverage obtained in metagenomic studies, and briefly presented the advantages of different approaches. Figure 4-2 shows how coverage is not simply a function of data set size (often the only indication to coverage in metagenomic studies), but largely depends on the complexity of the communities sampled. Figure 4-1 shows that quantitative estimations of coverage can serve as a basis for the adjustment of statistical tests, applicable to most, if not all, metagenomics studies. We recommend using at least one of the above-mentioned tools to estimate coverage (directly or indirectly) when analyzing metagenomic data sets, taking into consideration the objectives of the study and the nature of the data sets.

Acknowledgments

We thank Michael R Weigand for helpful suggestions regarding the manuscript. Our work is supported in part by the US DOE Office of Science, Biological and Environmental Research Division (BER), Genomic Science Program, Award Nos. DE-SC0006662 and DE-SC0004601, and by the US National Science Foundation Award No. 1241046.

CHAPTER 5: NONPAREIL: A REDUNDANCY-BASED APPROACH TO ASSESS THE LEVEL OF COVERAGE IN METAGENOMIC DATASETS

Originally published on March 2014 in *Bioinformatics* 30 (5): 629-635[1],
reproduced by permission of Oxford University Press,
DOI: 10.1093/bioinformatics/btt584.
Luis M. Rodriguez-R & Konstantinos T. Konstantinidis.

Motivation: Determining the fraction of the diversity within a microbial community sampled and the amount of sequencing required to cover the total diversity represent challenging issues for metagenomics studies. Owing to these limitations, central ecological questions with respect to the global distribution of microbes and the functional diversity of their communities cannot be robustly assessed.

Results: We introduce Nonpareil, a method to estimate and project coverage in metagenomes. Nonpareil does not rely on high-quality assemblies, operational taxonomic unit calling or comprehensive reference databases; thus, it is broadly applicable to metagenomic studies. Application of Nonpareil on available metagenomic datasets provided estimates on the relative complexity of soil, freshwater and human microbiome communities, and suggested that ~200 Gb of sequencing data are required for 95% abundance-weighted average coverage of the soil communities analyzed.

---

[1] Associate editor: Michael Brudno.

Availability and implementation: Nonpareil is available at *https://github.com/lmrodriguezr/nonpareil/* under the Artistic License 2.0.

Introduction

Metagenomics have provided important new insights into the diversity, dynamics and functional potential of natural microbial communities during the past decade, but several critical issues remain unresolved. Many metagenomic surveys to date have sampled only a small fraction of the total community DNA; this is particularly the case for soil and sediment communities. Furthermore, the amount of sequencing required to cover the whole community remains speculative (Wendl et al., 2012). The fraction of the genomes recovered in a sequencing dataset is termed coverage (Suppl. Box S1), and depends on the sequencing effort applied and the diversity of the community. When the coverage of a metagenome is unknown, results and conclusions about species richness, the evenness of the corresponding community, differences between communities and the extent and importance of rare community members are limited. Moreover, differences between sequencing technologies and continuously changing sequence read lengths make it challenging to establish a universal approach for coverage estimation.

Determining the total number of unique species or operational taxonomic units (OTUs) present in a sample is frequently challenging due to the unknown number of non-sampled species and requires either complete coverage or knowledge of the species abundance curve, which typically remains elusive. The coverage achieved by a dataset can be calculated more efficiently, as it does not depend on *a priori* knowledge of the species abundance curve, and can be directly related to assembly quality (Wendl, 2006). The level of coverage is typically assessed by identifying and counting OTUs and generating rarefaction curves (Hughes, Hellmann, Ricketts, & Bohannan, 2001). Empirical and analytical models have also been applied to coverage estimation using read binning

(Hooper et al., 2010; Stanhope, 2010) if assembly is not limiting or by targeting specific taxa (Wendl, 2006; Wendl et al., 2012) when genome size and abundance are known. However, these approaches and their variations (Hooper et al., 2010; Schloss & Handelsman, 2008; Stanhope, 2010; Tamames, de la Peña, & de Lorenzo, 2012; Wendl, 2006; Wendl et al., 2012) require either the use of a reference genome database or the clustering of reads in contigs or OTUs. The former is severely limited by the shortage of representative genome sequences from most habitats (D. Wu et al., 2009). The latter is limited by the quality of the assemblies, especially for highly complex communities, and the use of genes that are much more conserved than the genome average to be sufficiently similar to allow clustering of reads in OTUs such as the ribosomal RNA genes. These genes, however, are known to miss important levels of ecological differentiation among closely related, yet distinct, OTUs (Konstantinidis & Tiedje, 2007). Therefore, a method to estimate the coverage of a metagenomic dataset that is applicable to communities of varied diversity and does not depend on the quality of the assembly and the completeness of reference databases is highly desirable.

Here we introduce Nonpareil (‘having no match or equal’, referring to the count of unmatched reads in a dataset), a novel method that aims to fulfill this critical gap in contemporary metagenomic research. Nonpareil examines the degree of overlap among individual sequence reads of a whole-genome shotgun (WGS) metagenome to compute the fraction of reads with no match, which is used to estimate the abundance-weighted average coverage (i.e. not the arithmetic mean based on all species in the sample but the average when the abundance of species is considered). Subsequently, it fits a projection line to the estimated values to determine the amount of sequencing required for almost complete diversity coverage. The fraction of unmatched elements in a given subset of a finite collection (singletons in clustering terms) can be used to efficiently estimate the coverage of the collection, i.e. the fraction of the collection captured in the

subset (Esty, 1986; Good, 1953). This observation has been previously applied to metagenomic datasets to estimate species richness (Chao, 1984), functional coverage (Schloss & Handelsman, 2008) and coverage of gene amplicons (Schloss et al., 2009) based on ribosomal RNA or other individual genes. To the best of our knowledge, Nonpareil is the first method directly applying this concept to the whole-genome level, without using reference markers. Further, we propose that Nonpareil projection curves serve as a semi-quantitative proxy to the diversity of the communities. This feature is explored to rank natural communities in terms of the degree of their diversity.

Methods

Our method relies on the observation that datasets with higher coverage are more redundant because the sequencing reads are nearly random, although some systematic biases have been noted for specific sequencing protocols (Dohm, Lottaz, Borodina, & Himmelbauer, 2008). Redundancy is defined here as the portion of reads in a dataset that match with at least one other read (redundant reads; redundant portion is denoted κ). Calculating this value is computationally expensive because it requires a number of paired comparisons asymptotically equal to a quadratic growth (in the worst case, where no two reads match). This is a prohibitive calculation, even for powerful computers, for real-size sequencing datasets that are composed of millions of sequencing reads. Instead, Nonpareil estimates the redundancy value by generating a sample of query reads from the entire dataset (query subset), after which the number of matches per query read in the entire dataset is calculated. For each query read, the total number of matches in the complete dataset is calculated and stored (match-vector; Figure 5-1a). Based on the concept of the collector's curve, a saturation function of the redundancy is subsequently produced (Figure 5-1b), by iteratively sampling the match-vector in two steps. First, a subset of query reads is selected with a Bernoulli trial per read (with parameter equal to the sampling portion). Next, for each selected query read, the probability of matching another

read in the sample is estimated following a binomial distribution, i.e. the number of expected matches of the read in the sample decreases proportionally to the size of the sample, as described in Equation 5-1.

$$\Pr(m \geq 1) = 1 - \Pr(m = 0) = 1 - \binom{n}{0} p^0 (1-p)^n$$

$$\Pr(m \geq 1) = 1 - \left(1 - \frac{M-1}{N-1}\right)^{(N \times portion)-1}$$

Equation 5-1

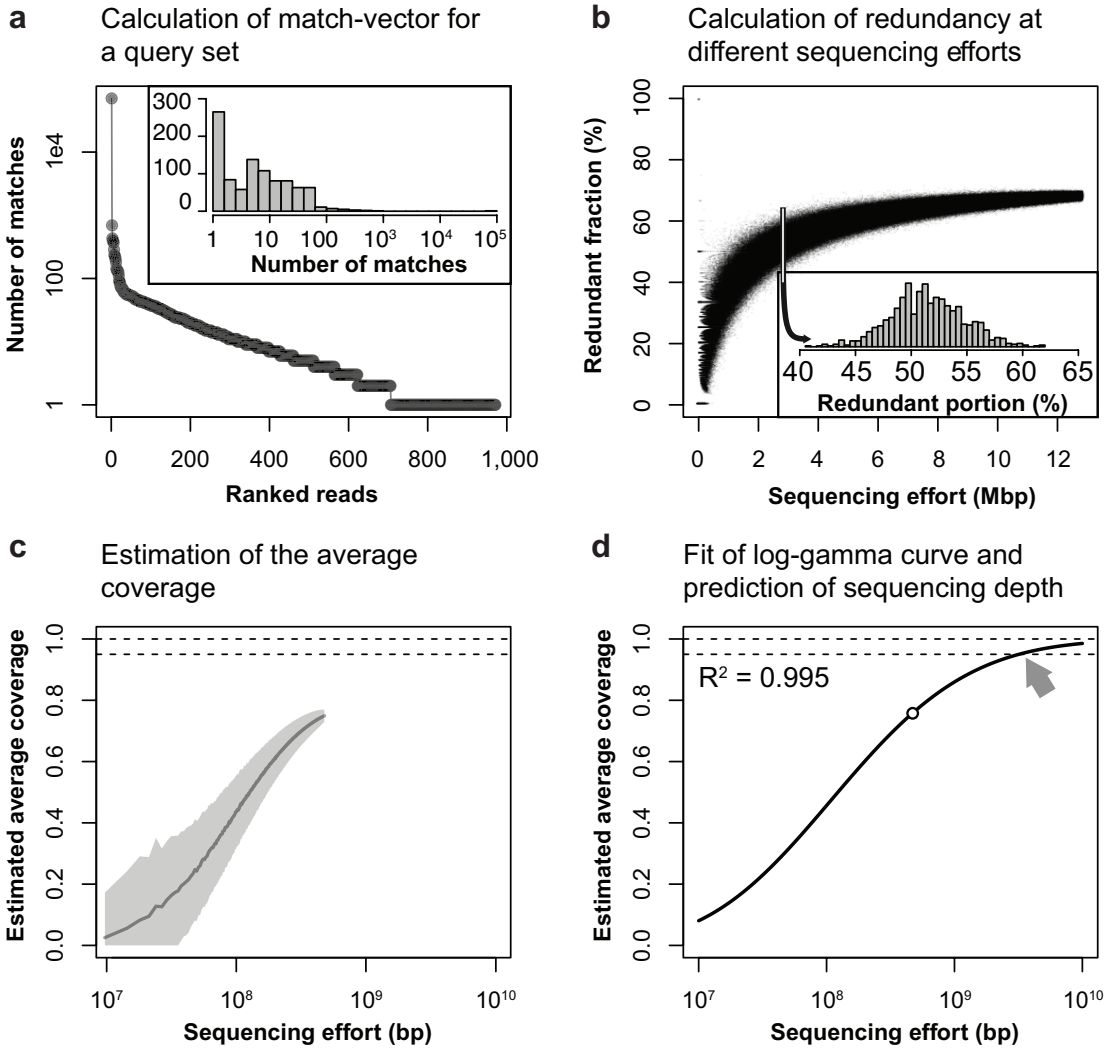


Figure 5-1. Main steps in the construction of Nonpareil curves.

A. The construction of a Nonpareil curve starts with the calculation of a vector containing the number of matches for a randomly drawn query subset from the total dataset (1,000 reads in this case). The function of the number of matches for each query read, ranked by decreasing number of matches, resembles a rank-abundance plot. The inset shows the histogram of matches for the same vector, *i.e.*, the observed distribution of matches from which the rank-abundance plot is generated. **B.** Next, the redundant portion is calculated for sub-datasets of different sizes. For each size, 1,024 replicate datasets are generated. The inset shows the distribution of replicates for a given dataset size. **C.** The distribution of redundancies is summarized by the average on each size, and the average coverage is estimated. The sequencing effort is displayed in logarithmic scale. The shadowed area represents one standard deviation from the average of the distribution. **D.** Finally, the curve of estimated coverage is fitted to a log-gamma function and is projected to predict the sequencing effort required to reach a given level of coverage. The solid line represents the fitted function; the empty circle indicates the size of the dataset; and the horizontal dashed lines indicate 100 and 95% coverage. The gray arrow indicates the point where the fitted Nonpareil curve reaches 95% average coverage**.**

Where *m* is the number of redundant reads in the subsample, *n* is the number of reads in the sample, *p* is the probability of finding a redundant read in the entire dataset, *M* is the number of redundant reads in the entire dataset, *N* is the total number of reads in the entire dataset and *portion* is the sampling portion of the entire dataset used for the estimation. This technique prevents redundant comparisons between reads because all comparisons are precomputed once, allowing the calculation of a Nonpareil curve with high resolution (i.e. with sampling portions close to each other). More importantly, it allows multiple replications at each sampling portion (1,024 times by default), reducing the effect of randomness in sampling. The resulting function is next summarized (calculating the average, median and standard deviation at each sequencing effort) to estimate the average coverage (Figure 5-1c). Finally, a log-gamma regression is fit to the calculated redundancy values with the weighted NL2SOL algorithm (Dennis, Gay, & Walsh, 1981) (Figure 5-1d). The projected regression

line allows for calculation of the sequencing effort required to reach a fixed average coverage. Nonpareil includes additional optimizations to further decrease the running time and required resources for estimating the redundancy values of datasets of several million of reads on a personal computer, as described later in the text.

Pairwise read comparison

Nonpareil performs ungapped alignments between reads using a sliding sequence approach that is included in the source code. The alignment strategy aims to match a prefix of the first sequence to a suffix of the second and vice versa. The search space is constrained by the minimum read overlap, excluding comparisons too short to satisfy the threshold, and by minimum identity (default is 95% nucleotide identity; see later in the text), discarding comparisons once the number of allowed mismatches is exceeded. Two sequences are considered redundant (matching pair) if they have at least one alignment satisfying both the minimum overlap and the minimum identity in any strand orientation. Ungapped alignments were preferred because the search space is much smaller than that of gapped alignments, with significant improvements in computation, whereas insertions and deletions between highly identical sequences (i.e. 95% nucleotide identity or higher) occur at a low frequency. Further, the new sequencers such as the Illumina MiSeq and HiSeq platforms show low rates of insertion and deletion sequencing errors, *i.e.*, <0.01% (Dohm et al., 2008). Finally, sliding sequences are meant to detect overlapping sequencing reads originating from the same genomic region (Wendl, 2006), as opposed to reads with high overall similarity (global alignments) or sharing regions that are not necessarily terminal (local alignments).

Simulated datasets used in this study

To benchmark our method and resolve the numeric relationship between the redundancy value (κ) calculated by Nonpareil and the average coverage ($\bar{C}$) of a dataset, we generated 120 training datasets by sampling publicly available complete bacterial and archaeal genomes from NCBI's GenBank database uniformly at random (independently of their length and nature). For each dataset, we selected a variable number of genomes, ranging from 1 to 1262. We generated 13 additional datasets using only 282 genomes from *Escherichia coli*, *Yersinia pestis*, *Helicobacter pylori* and *Staphylococcus aureus* to simulate environments with low species richness and phylogenetic diversity, but high intra-species diversity (termed hereafter 'Low richness'). Finally, we produced 10 datasets using 130 genomes from the genus *Escherichia*, simulating environments with extremely low phylogenetic diversity (termed '*Escherichia*'). We randomly assigned an abundance value to each genome in the sample from an exponential distribution ($\lambda = 1$) and produced a number of reads from each genome relative to that value. To produce Illumina-like reads, we randomly selected positions in the genome from a uniform distribution and generated a 101 bp-long read from either strand with randomly introduced sequencing errors from a binomial distribution ($P = 0.01$, $n = 101$). We used the resulting training datasets to evaluate the correlations between Nonpareil indices (κ and R^*) and estimate the average coverage and required sequencing effort for nearly complete coverage (Suppl. Table S1) in log-log space.

Sequencing depth and coverage estimation

To estimate the coverage of a genome within a training metagenome (generated *in silico*), we backtracked the reads generated from the genome, regardless of the amount of error or the orientation of the reads, and divided the number of covered positions by the genome length. Note that coverage (C) and sequencing depth (ρ; see Suppl. Box S1) share a close relationship, generally approximated

through the Lander–Waterman equations (Lander & Waterman, 1988), Equation 5-2.

$$C = 1 - e^{-\left(\frac{LR\alpha}{\gamma}\right)} = 1 - e^{-\rho}$$

Equation 5-2

Where LR is the average read length times the number of reads (*i.e.*, the sequencing effort), α is the abundance of the target genome, γ is the length of the genome and ρ is the sequencing depth. To estimate the sequencing depth of a given genome [ρ; Equation 5-3], we simply divided the added length of the reads originating from the genome (T) by its length (γ). Note that Equation 5-2 implies that the number of reads (T) equals the abundance times the number of reads in the dataset (αR).

$$\rho = \frac{TL}{\gamma}$$

Equation 5-3

Accordingly, we defined average sequencing coverage of a sample ($\bar{C}$) as the sum of the sequencing coverage of each genome (C_i) multiplied by its sequence probability [π_i; Equation 5-4].

$$\bar{\gamma} = \sum_i \alpha_i \gamma_i \qquad \pi_i = \frac{\alpha_i \gamma_i}{\bar{\gamma}} \qquad \bar{C} = \sum_i \pi_i C_i$$

Equation 5-4

For simplicity, this estimation does not take into account non-observed species because no assumptions about the distribution of abundances of those can be made without additional information. However, species under the detection level are expected to only marginally affect the estimation of both the average

coverage and the average genome size because the contribution of each species (i) depends on its abundance (α_i). In cases where the coverage is extremely low, the contribution of non-observed species to the total community can be relatively high, causing unreliable estimates. Nonpareil automatically identifies datasets with estimated coverage below 10^{-5} or with median redundancy of zero in 20% of the subsample and reports them as insufficient data. Although available approximations might provide marginally better estimations of the average sequencing depth (Hooper et al., 2010; Tamames et al., 2012), they rely on assumptions about the shape of the distribution of abundance, which may be unrealistic.

Estimation of sequencing efforts for nearly complete coverage

To estimate the amount of reads needed to attain a nearly complete coverage of a simulated community/sample, we used Equation 5-1 to estimate the number of reads necessary to cover 95% of every target genome [R^*_i in Equation 5-5] and calculate the average of these values [R^* in Equation 5-6]. Note that the value of R for which C equals one is undefined [Equation 5-2], and we use $C = 0.95$ as a rule-of-thumb for *nearly complete coverage* [(Bouck, Miller, Gorrell, Muzny, & Gibbs, 1998); *cf.* (Wendl et al., 2012) for discussion].

$$R_i^* = \frac{-\ln(1-0.95)\gamma_i}{\alpha_i L} \approx 3\gamma_i/\alpha_i L$$

Equation 5-5

$$R^* = \sum_i \pi_i R_i^*$$

Equation 5-6

We define here the sequencing effort required for *nearly complete coverage* of a community (*R**) as the expected number of reads necessary to produce an average coverage of at least 95% of the genomes in all sampled cells.

Nonpareil curve construction and model fitting

For any given dataset, the Nonpareil curve is defined as the average coverage (estimated from the portion of reads that is similar to at least one other read in the sample; κ) as a function of the sample size (*LR*). Two reads are assumed to be similar if their ungapped alignment shows similarity and length coverage above user-defined thresholds. Here, we used 95% nucleotide sequence identity, intended to reflect natural discrete populations and current species demarcation standards (Caro-Quintero & Konstantinidis, 2012; Goris et al., 2007) and exceed typical sequencing error (Dohm et al., 2008); and three values of alignment length: 25, 50 and 75% of the length of the shortest read. Although we observed that 50% overlap is generally optimal for most datasets, comparisons with 25% overlap should be preferred in extremely low coverage datasets that may be challenging to analyze with 50% overlap. On the other hand, comparisons with 75% overlap may produce fast preliminary results, suitable for high coverage datasets (the longer the overlap, the faster the computation of κ).

The Nonpareil curve has a dual purpose. First, it shows the portion of redundant reads in the entire dataset, reflecting the coverage of the dataset. Second, it allows a projection from the data in hand to the sequencing effort required to achieve any user-defined portion of redundancy, reflecting the complexity of the sample. To perform the projection, the Nonpareil curve is fitted to the cumulative probability function of the gamma distribution [with log-transformed values of the sequencing effort *R*; Equation 5-7].

$$\kappa = \frac{\gamma(a, b \cdot log(R+1))}{\Gamma(a)} = \frac{\int_0^{b \cdot \log(R+1)} e^{-t} t^{a-1} dt}{\int_0^{\infty} e^{-t} t^{a-1} dt}$$

Equation 5-7

Where Γ is the gamma function, γ is the lower incomplete gamma function (both explicitly noted in the rightmost part of the expression), κ is the redundant portion (coordinate axis in the curve), R is the sample size (ordinate axis in the curve) and a and b are parameters that determine the shape and rate of the curve, respectively, estimated using the weighted NL2SOL algorithm (Dennis et al., 1981).

Implementation

We implemented the Nonpareil algorithm in C++ with an ancillary R script for model fitting and plotting, using only standard C++ and R libraries. The software parallelizes the read comparisons and sampling steps with an arbitrary number of threads. To reduce the number of hard drive access requests without compromising memory efficiency, blocks of reads are loaded into memory with a maximum random-access memory (RAM) usage defined by the user. This allows the software to run with modest minimal requirements, while ensuring scalability in high-performing computers.

Real metagenomic datasets

We generated Nonpareil curves for a collection of metagenomic datasets from different environments and levels of diversity. In all cases, we used Nonpareil with default parameters: sequence identity of 95% and read overlap of 50%. We considered the acid mine drainage (AMD) dataset as an example of a community with extremely low phylogenetic diversity (Denef & Banfield, 2012). The sample from site C75 (July 2011), was composed of only *Leptospirillum* sp. group II genotype III, and 5% and 1% subsets of this sample were used to calculate the Nonpareil curves. The genome length of *Leptospirillum* sp. was assumed to be 2.6 megabase (Mb; added length of the scaffolds from GenBank entry

AIJM00000000) to calculate both the expected coverage and the expected number of reads required to achieve nearly complete coverage.

We analyzed six selected datasets from the Human Microbiome Project, or HMP (The Human Microbiome Project Consortium, 2012), for which both WGS and amplified 16S ribosomal RNA gene (16S) sequencing data were available (Suppl. Table S2). To compare Nonpareil results with a 16S-based estimation, we employed COVER with default parameters (Tamames et al., 2012) to predict the abundance (corrected by 16S copy number) and the genome size of the OTUs in the community from the 16S amplicon data, and used this information to calculate the average sequencing depth and the required effort for nearly complete coverage of the community (Suppl. Table S2). COVER reports the sequencing effort required to achieve a given coverage or a given sequencing depth in the top-n OTUs, but we employed the estimation of R^* on Equation 5-6 (based on abundance and genome length predicted by COVER) to allow comparisons with our method. We also used OTU tables (Caporaso et al., 2010) based on 16S data from http://www.hmpdacc.org/ (The Human Microbiome Project Consortium, 2012) to independently assess abundance distributions and Chao1 indexes (Chao, 1984).

We calculated Nonpareil curves for datasets from Lake Lanier (GA, USA), Hess Creek (AK, USA), and the Manu National Park (Peru), representing complex natural environments. We included two samples from August 2009 from Lake Lanier (LL-S1 and LL-S2; (Oh et al., 2011)), and one additional sample from the same site from July 2010 (LL_1007B) with over twice the sequencing effort, generated with Illumina GA II (100 bp paired-end reads); two soil samples from Hess Creek representing the active and the permafrost layer of the core 2 of day 2 (Mackelprang et al., 2011); and one soil sample from the tropical forest of Manu National Park in Peru (PE6; (Fierer et al., 2012)). All datasets were trimmed using SolexaQA (Cox, Peterson, & Biggs, 2010) with maximum expected error of

1% and minimum length of 50 bp. In paired-end samples, only the forward reads were used.

Results

Nonpareil curves were calculated for 143 short-read simulated metagenomes of various size and diversity levels, generated from publicly available bacterial genomes. Nonpareil estimates of the average coverage of each metagenome correlated strongly (Pearson's R^2 = 0.93, n = 126) with the independently calculated coverage values based on the known composition of each metagenome. Further, the amount of sequencing that was required to nearly cover the total diversity predicted by Nonpareil (abundance-weighted average coverage of 95% for the genomes in the sample) corresponded tightly to the actual values for each metagenome (Suppl. Fig. S1 and S2, and Suppl. Table S1; Pearson's R^2 > 0.65). The estimated abundance-weighted average coverage may also serve as an indicator for the expected quality of the metagenome assembly. Although several factors other than the coverage are critical for assembly, our results show that the average coverage provides a lower-bound estimation of the fraction of assembled reads (Suppl. Fig. S3a), while the assembly N50 of samples with coverage below 60% rarely surpasses twice the read length for Illumina datasets (Suppl. Fig. S3b).

It is important to note that the precision of the algorithm was reduced at values of redundancy (κ) lower than 1% and higher than 90%. These values approximately correspond to <0.01X and >400X sequencing depth, respectively (Suppl. Fig. S4). Because datasets with lower sequencing depth than 0.01X (i.e. too few sequences obtained) are strongly influenced by random variability and thereby subject to spurious results, Nonpareil estimates are not reliable at this range and such datasets are flagged accordingly in the output of the algorithm. Conversely, datasets with sequencing depth above maximum saturation (>400X) are best

assessed by read recruitment (mapping), as high-quality assemblies should be achievable in these cases.

Influence of sequencing error

High frequency of sequencing errors can affect the estimations of the number of redundant reads and thus, Nonpareil curves. It is strongly recommended to filter reads with a stringent cutoff for expected error (e.g. resulting in reads with <1% error rate) prior to applying Nonpareil. The distribution of sequencing error is not always uniform across the length of the reads, depending on the sequencing platform used, and this uneven distribution may affect Nonpareil estimates. In order to evaluate the latter, we analyzed a ~61 Mb dataset generated *in silico* from 33 reference genomes, dominated by *Aeromonas salmonicida* subsp. salmonicida (14%), in which randomly introduced sequencing errors were distributed uniformly, increasing linearly, and increasing as a polynomial of order 4 across the read length (based on (Korbel et al., 2009)). These involved only wrong-base substitution errors, the dominant source of error in Illumina. The resulting curves (Suppl. Fig. S5) indicated that the estimates of Nonpareil are not affected by the distribution of errors when the total error is ~1% or less, but can be strongly biased when sequencing error approaches 5%. For other types of sequencing errors such as artificial duplicates, a common artifact in 454 sequencing, it is recommended to detect and remove sequences with these errors (e.g., (Balzer, Malde, Grohme, & Jonassen, 2013)) prior to applying Nonpareil.

Coverage estimation of various natural communities

Application of Nonpareil curves to publicly available metagenomes revealed, as expected, that the soil samples required the highest sequencing effort for nearly complete coverage. The seasonally thawed active soil layer from a black-spruce forest in the discontinuous permafrost zone of Alaska at Hess Creek required the

highest sequencing effort (Table 5-1, 0.2 Tb) for nearly complete coverage. The freshwater samples were predicted to require ~10 times less sequencing than soil but more sequencing compared with all evaluated human microbiota (e.g., >80X more than posterior fornix; Figure 5-2 and Table 5-1). The AMD sequences mapping to *Leptospirillum* sp. covered 99.99% of the genome, and the 1% subset covered 73%. Using Nonpareil, a sequencing coverage of 70% was estimated for the 1% subset, which corresponds to 17 Mb of the complete dataset, whereas an expected coverage of 94% was obtained using the Lander-Waterman expression [(Lander & Waterman, 1988); Equation 5-2].

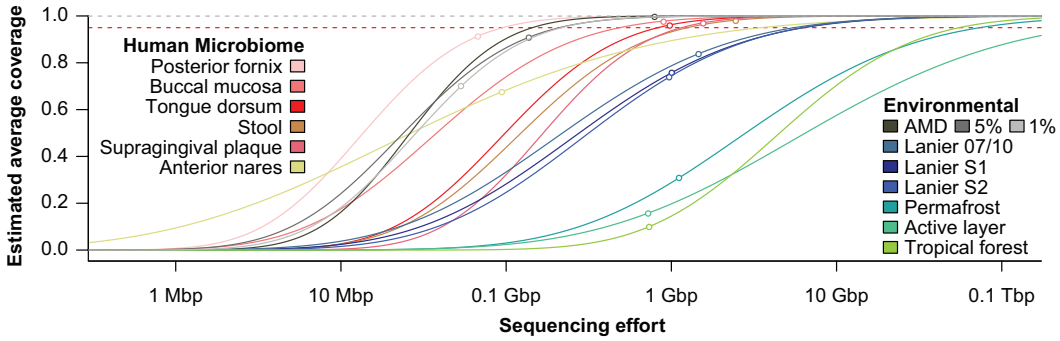


Figure 5-2. Comparison of Nonpareil curves for the metagenomes of HMP, AMD, Lake Lanier, Permafrost soil and Tropical Forest soil.

The plot displays the fitted models of the Nonpareil curves. The horizontal dashed lines indicate 100 (gray) and 95% (red) coverage. The empty circles indicate the size and estimated average coverage of the datasets, and the lines after that point are projections of the fitted model. The curves cluster in four groups, reflecting different levels of diversity. The leftmost group, including posterior fornix, buccal mucosa, anterior nares and AMD, represents samples largely dominated by a single species. A second group, composed by tongue dorsum, stool and supragingival plaque, represents low to medium diversity samples. Next, freshwater samples, which are typically characterized by moderate to high diversity, cluster together. Finally, the curves for the high-diversity soil samples display the lowest slopes.

In addition to WGS metagenomes, the HMP samples included amplified 16S ribosomal RNA gene (16S) sequencing data (Suppl. Table S2). Estimation of the abundance and genome size of 16S-defined OTUs by COVER (Tamames et al., 2012) displayed larger variability (*cf*. Table 5-1 and Suppl. Table S2) compared with Nonpareil estimates based on metagenomes from the same samples, possibly reflecting the influence of sequencing errors or polymerase chain reaction artifacts (Kunin, Engelbrektson, Ochman, & Hugenholtz, 2010) or variations on the assumed number of 16S copies (Větrovský & Baldrian, 2013). The largest difference between COVER and Nonpareil was observed in the posterior fornix, an environment known to be largely dominated by *Lactobacillus* (Ravel et al., 2011). Quality-checked 16S data showed that the most abundant OTU accounted for ~90% of the community and only nine of 37 OTUs identified showed abundance >0.1%. Assuming a typical vaginal lactobacilli genome size of 2.4 Mb, ~8 Mb are predicted by the Lander-Waterman expressions (Lander & Waterman, 1988) to be required to cover 95% of the dominant OTU. Nonpareil estimated that 12 Mb of sequence data provide an average coverage of 91%, and ~40 Mb would be necessary to cover the community almost entirely. In contrast, COVER estimated a total of 226 OTUs and 160 Gb to be necessary for the same level of coverage. These results suggest that 16S analysis can frequently inflate diversity, resulting in large underestimations of the sequencing depth. They also reveal that Nonpareil produces estimations closer to those of other genome-wide approximations such as the Lander–Waterman model. We limited our evaluation to COVER because alternative methods for coverage estimation (Daley & Smith, 2013; Hooper et al., 2010; Stanhope, 2010; Wendl et al., 2012) were either not available for online or standalone computation, do not scale with large metagenomic datasets or provide results like longest contig expected per taxon that are not directly comparable with those of Nonpareil.

Table 5-1. Nonpareil estimates for publicly available metagenomic datasets.

All analyses were executed with default parameters (50% read overlap, 95% identity). For each sample, Nonpareil estimated the average coverage ($\hat{C}$) and predicted the sequencing effort required for nearly complete coverage (LR^*). Dashes (-) indicate that the model was not projected because the estimated coverage exceeds 95%. Identifiers starting with SRS indicate entries in the NCBI Sequence Read Archive; all other identifiers are from to the original publications.

Location	Identifier	LR	$\hat{C}$	LR*	Data source
Anterior nares	SRS019087	20 Mb	68%	2.3 Gb	(The Human Microbiome Project Consortium, 2012)
Buccal mucosa	SRS063287	1.0 Gb	95%	-	
Stool	SRS016335	5.6 Gb	97%	-	
Supragingival	SRS015574	2.5 Gb	97%	-	
Tongue dorsum	SRS062540	1.2 Gb	95%	-	
Posterior fornix	SRS063417	12 Mb	91%	43 Mb	
Richmond mine (CA, USA)	C751107	0.7 Gb	99%	-	(Denef & Banfield, 2012)
	C751107 1%	7.2 Mb	70%	90 Mb	
Lake Lanier (GA, USA)	LL-S1	1.1 Gb	72%	3.2 Gb	(Oh et al., 2011)
	LL-S2	1.1 Gb	73%	3.1 Gb	
	LL_1007B	2.3 Gb	83%	2.8 Gb	This study
Hess Creek (AK, USA)	Permafrost C2	1.8 Gb	31%	41 Gb	(Mackelprang et al., 2011)
	Active C2	1.6 Gb	16%	198 Gb	
Manu Park (Peru)	PE6	0.7 Gb	10%	30 Gb	(Fierer et al., 2012)

Diversity ranking

An interesting feature of the Nonpareil curves is that the shape of the curves reflects the level of diversity of the communities sampled. The Nonpareil curve saturates faster, *i.e.*, complete coverage is achieved with fewer sequences sampled, on datasets with lower diversity and shorter genomes (Figure 5-2). Because the average genome sizes differ by no more than one order of

magnitude between most microbial communities, the velocity of saturation of Nonpareil curves is mostly determined by the sample diversity rather than differences in genome size or gene duplications and repetitive regions. However, deviations from this expectation are possible when comparing metagenomes with large differences in average genome size, as it is often the case when the proportions of viral, bacterial/archaeal and eukaryotic DNA differ substantially. In such cases, separation of the different fractions (*e.g.*, (Liu et al., 2013)) before applying Nonpareil is recommended and the efficiency of this technique needs to be assessed on a case by case basis. Figure 5-2 revealed clustering of curves from samples with decreasing levels of diversity. Nonpareil curves from samples of communities characterized by low diversity, like posterior fornix, anterior nares and AMD, rapidly saturated. In contrast, Nonpareil curves from soil samples, known to possess comparatively high diversity, continued growing after projecting to millions of reads. Intermediate in Figure 2 are Nonpareil curves from freshwater samples, stool and tongue dorsum, expected to have a higher diversity than the first group of samples but lower than soil datasets. This property of the Nonpareil curves allows fast assessment of the level of diversity inherent to an unknown sample compared with reference communities. In addition, the shape of the Nonpareil curves can reveal distinctive features of the samples such as skewed distribution of species abundances. For example, the Nonpareil curve for the anterior nares sample (Figure 5-2) showed a rapid growth phase at low sequencing effort that does not saturate as rapidly as other low-complexity samples. Further examination indicated that this pattern was due to an unusual distribution of abundances (as revealed by 16S profiling; (The Human Microbiome Project Consortium, 2012)), following an extreme broken-stick model. In all, 74 species were observed and ~99 species were estimated to coexist in this sample (Chao1, $IC_{95\%}$: 32.98–239.7) but the most abundant species had an abundance of 36%, and the 9 most abundant species represented 95% of the community.

Computing performance

We tested Nonpareil with datasets of various sizes (101 bp-long reads) and evaluated its performance in terms of central processing unit (CPU) time, running time and RAM usage (Suppl. Fig. S6). All tests were performed on cluster architecture with 64 CPUs (2.2 GHz) per node, >40 GB of available RAM, running on Red Hat Enterprise Linux 6. Both the running time and the RAM usage grow linearly with the size of the dataset, as anticipated. The RAM use in GB was ~0.1 times the size of the dataset (in millions of reads) plus 2. This relationship might vary on different computers, operating systems and future versions of the code, but it offers an indication of the RAM requirement of the algorithm without parceling. Note that the maximum RAM usage can be set on each run by the user, and Nonpareil can parcel the data to adapt to less powerful computers as needed. Both the running time and the CPU time are strongly affected by the stringency (cutoffs) of the read comparison (Suppl. Fig. S6 B and C). However, the algorithm scaled up equally well with all the parameters (Suppl. Fig. S6D).

Conclusions

The results presented here highlight the usefulness of the Nonpareil curve as a tool for both study design and exploratory comparisons of community diversity. This tool increases the range of samples for which coverage can be computed relative to existing tools. It is important to point out that existing approaches for coverage estimation require prior knowledge about the abundance distribution of the members of the community (Wendl, 2006; Wendl et al., 2012) and/or assume that the diversity distribution can be effectively modeled by known probability distributions (Hooper et al., 2010; Stanhope, 2010), or require the use of reference molecular markers (Daley & Smith, 2013; Tamames et al., 2012). These properties of a metagenome are frequently not available. The relationship between sequencing effort and average coverage of the community can be

alternatively approximated by visual inspection of rarefaction when binning is feasible (Schloss et al., 2009; Schloss & Handelsman, 2008). A recent development improved on this traditional approach by providing a mathematical generalization for any molecular marker and an accurate projection of the rarefaction curve (Daley & Smith, 2013). However, the level of coverage remains inaccessible and sequence binning is a required step, which is typically limiting in WGS metagenomic studies. In contrast, Nonpareil does not require abundance distributions, models or reference databases and is based on the redundancy of the reads, an intrinsic characteristic of any metagenomic dataset. The complement of redundancy is the number of reads without matches in a given sample divided by the sample size, which we denoted as the Nonpareil fraction (u). When expressed in terms of non-matching reads (i.e. one minus the Nonpareil fraction) the redundancy essentially takes the same form as the Good's coverage estimator (Good, 1953), a widely applied estimator of coverage of a sample (Esty, 1986). Nonpareil applies this estimation directly on shotgun sequencing reads, even in datasets composed of millions of reads, with modest computational requirements.

Application of Nonpareil estimates on available metagenomes revealed, as expected, that the largest sequence efforts were required for soil datasets, where up to 200 Gb and 1 Tb of sequence data were predicted to be necessary to achieve 95 and 99% abundance-weighted average coverage, respectively. These estimates are well below the 10 Tb estimate of (Riesenfeld, Schloss, & Handelsman, 2004) required to cover a typical soil metagenome, which emphasizes on coverage of all species, including rare ones. For example, Nonpareil predicts an increase in average coverage from 99.9 to 99.99% with 1–10 Tb of data (in Hess Creek), a marginal difference in abundance-weighted average coverage for 10 times more data. These results agree with previous findings based on single target species (Wendl et al., 2012), supporting that the estimations of Nonpareil are practical and robust. The soil dataset of Hess Creek

represents a permafrost soil incubated under warm temperatures, which likely stimulated specific taxa, affecting the diversity of the community. However, the Manu Park sample represents a temperate soil, estimated to contain close to 9000 species based on 16S data (from 5347 observed 97% OTUs; (Fierer et al., 2012)). In fact, the estimate provided by Nonpareil for 99% average coverage (95 Gb) translates to complete coverage of any genome of ~5 Mb with abundance >0.07 with 90% confidence (Wendl et al., 2012). This corresponds to the top 312 most abundant species, or 90% of the observed community, based on 16S (Fierer et al., 2012). Note that these 312 species likely represent only ~5% of the number of species present in the community. However, Nonpareil estimate is not meant to reflect the captured richness of the community (*i.e.*, how many different species were captured), but the portion of the total community captured, taking abundance into consideration.

Finally, we evaluated the robustness of Nonpareil estimates by both decreasing the sequencing effort on a community with high coverage (AMD) and increasing it on a community with medium coverage (Lake Lanier). In both cases the estimates were consistent with the expectations (Figure 5-2 and Table 5-1), indicating that Nonpareil analysis is robust to variations in the size of the query dataset, and variations arising from independently collected samples or different sequencing protocols (Lake Lanier samples).

In summary, Nonpareil curves offer an estimation of average coverage of metagenomic datasets (for profiling studies and other community-wide analyses), a prediction of coverage in increased sequencing efforts (for study design), and a comparative framework for diversity exploration, allowing for fast diversity rankings of metagenomes before assembly or taxonomic classification.

Availability

Nonpareil is free software licensed under the terms of the Artistic license 2.0. The source code and binaries are available at https://github.com/lmrodriguezr/nonpareil/. An online server is available at http://enve-omics.ce.gatech.edu/nonpareil/. Sequences of Lake Lanier (LL_1007B) were deposited in the NCBI Sequence Read Archive, with accession number SRR948155.

Supplementary data

Supplementary data are available at http://bioinformatics.oxfordjournals.org/content/suppl/2013/10/10/btt584.DC1.

Acknowledgments

The authors thank Heidi Kizer, Janet Hatt, and three anonymous reviewers for helpful suggestions regarding the manuscript.

Funding

U.S. Department of Energy (Award DE-SC0006662) and by U.S. National Science Foundation (Award No 1241046).

CHAPTER 6: NONPAREIL 3: FAST ESTIMATION OF METAGENOMIC COVERAGE AND SEQUENCE DIVERSITY

Reproduced with permission from **Luis M. Rodriguez-R, Santosh Gunturu[1], Jiarong Guo[1], Chengwei Luo[2], James M. Tiedje[1,3,4], James R. Cole[1,4] & Konstantinos T. Konstantinidis**. All copyright interests will be exclusively transferred to the publisher upon submission.

Motivation: Estimations of microbial community diversity based on metagenomic datasets are affected, to an unknown degree, by biases derived from insufficient coverage and reference database-dependent estimations of diversity, while quantifying these biases is prohibitively time-consuming for large datasets and different sequencing technologies.

---

[1] Center for Microbial Ecology, Michigan State University, East Lansing, MI, USA.

[2] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

[3] Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA.

[4] Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, USA.

Results: We extended Nonpareil, a tool for the estimation of coverage in metagenomic datasets, to a high-performance computing implementation that scales up to hundreds of cores and includes, in addition, a k-mer based estimation as sensitive as but about three hundred times faster than the original alignment-based one. Further, we propose a metric of sequence diversity (N_d) derived directly from Nonpareil curves, which correlates well with alpha-diversity.

Availability: Nonpareil 3 is freely available at *https://github.com/lmrodriguezr/nonpareil*.

Introduction

The exploration of microbial diversity in natural and engineered environments has been revolutionized by the use of metagenomics. However, the power of both descriptive and comparative analyses in metagenomics is strongly deterred by low coverage, understood as the fraction of the DNA space covered by sequencing (Rodriguez-R & Konstantinidis, 2014b). To date, most metagenomics studies assess the level of coverage only indirectly or not at all, mainly owing to the difficulty in directly measuring the unseen fraction of a community. We have recently presented a method to assess the level of coverage in metagenomic datasets using a redundancy-based approach: Nonpareil (Rodriguez-R & Konstantinidis, 2014c). While we demonstrated that Nonpareil accurately estimates the level of coverage in metagenomic datasets, the estimation remained prohibitively expensive for datasets comprising several millions of base pairs. Here, we present a new version of Nonpareil that effectively distributes the estimation across processors and computing nodes, using High Performance Computing (HPC) capabilities now widely available, as well as an alternative estimation based on k-mer redundancy with comparable accuracy. Moreover, we previously showcased the qualitative sequence-diversity ranking derived from Nonpareil curves (Rodriguez-R & Konstantinidis, 2014b, 2014c). Here, we

quantitatively assess the level of sequence diversity derived from Nonpareil curves, and compare these estimations with other diversity indices.

Implementation

The processing of a sample in Nonpareil is divided into two main steps. The first step is the redundancy estimation, where sequences are compared and the estimated redundancy is subsampled at different values of sequencing effort. The large number of pairwise alignments makes this step the most resource-intensive. This task is now distributed across multiple nodes (Message Passing Interface; MPI) and CPUs (C++ pthreads). Further, Nonpareil 3 offers a k-mer-based method for redundancy estimation (Suppl. data §A.1.1) that accelerates this step by using short fragments of the sequencing reads with no errors allowed. Finally, redundancies are subsampled in multiple threads. In Nonpareil 1 subsamples were estimated linearly, resulting in sparse values towards the left side of the Nonpareil curve. While this strategy is still available, the default in Nonpareil 3 is logarithmic subsampling: sample size is iteratively multiplied by a density factor (default 0.7) until two reads remain.

The second step is the estimation of abundance-weighted average coverage at different sequencing efforts (Nonpareil curves), fitting to a sigmoidal model (projection), and graphical representation (Suppl. Fig. 1). This step has modest resource requirements and is implemented in the Nonpareil R package. In Nonpareil 3, we have streamlined this analysis using headers in the redundancy output files and included an estimation of the sequence diversity derived from the fitted model (*cf.* § Nonpareil Index of Sequence Diversity).

Reducing run time for large metagenomes

Nonpareil 3 can use parallelization across nodes and/or threads, resulting in time reductions of up to 500 times (Suppl. data §A.2.2). In addition, the read redundancy can now be estimated using perfect matches of one k-mer per query

read corrected by sequencing error estimation, instead of the complete ungapped alignment. This implementation results in similar coverage and diversity estimates, as well as highly correlated projections of sequencing effort, but reduces the computing effort necessary by about 300 times (Table 6-1, Suppl. data §A.2.1).

Table 6-1. Kernel comparison of Nonpareil estimates for publicly available datasets.

The two kernels (A: alignment, K: k-mer) where compared in terms of CPU time (*estimated for Iowa Soil, observed in all other cases), estimated coverage, and projected required sequencing effort to reach 95% coverage in samples from HMP (posterior fornix, tongue, stool), freshwater (Lake Lanier), and soil (Iowa continuous corn field).

Sample	Size (Gbp)	CPU Time (m)		Coverage (%)		Req. effort (Gbp)	
		A	K	A	K	A	K
P fornix	0.01	15.7	0.08	88	82	0.028	0.027
Tongue	0.22	286	0.68	67	59	1.09	1.21
Stool	0.32	438	0.85	79	69	1.04	1.39
LL 2011	2.95	4,397	16.5	83	77	6.78	8.79
LL 2009A	1.17	1,444	6.40	61	61	6.46	6.18
LL 2009B	1.12	1,463	5.75	69	61	4.97	5.36
Iowa soil	14.5	22,806*	49.0	53	44	137	149

Nonpareil Index of Sequence Diversity

Nonpareil curves are plots of abundance-weighted average coverage ($\hat{C}$) per sequencing effort (LR), fitted to the cumulative probability function of the gamma distribution (Rodriguez-R & Konstantinidis, 2014c) with parameters a and b:

$\hat{C}=\gamma(a, b \cdot \log(LR+1))/\Gamma(a)$, where Γ is the gamma function, and γ is the lower incomplete gamma function. Hence, we can use the mode of the corresponding gamma distribution to identify the value of *log(LR+1)* corresponding to the inflection point of the curve, which we propose as a measurement of sequence diversity: $N_d = (a-1)/b$. This index, with units of logarithm of base pairs, summarizes the community diversity in sequence space. Since the shape of the Nonpareil curves from replicates and subsamples closely resemble each other regardless of coverage (Rodriguez-R & Konstantinidis, 2014c), we propose N_d as a coverage-independent measurement of diversity for the source community. This metric depends on the joint distribution of genome size and abundance, as well as intra-genome gene duplication, and given a small variation in genome sizes and a small impact of genomic duplications (*e.g.*, for prokaryotic-only communities), N_d can be used as a database-independent metric of diversity. We compared Shannon diversity indices (H') evaluated on operational taxonomic units derived from 16S ribosomal RNA gene amplicon datasets (16S OTUs) from the Human Microbiome Project (HMP; (The Human Microbiome Project Consortium, 2012)) against N_d of metagenomes, and observed a high correlation (Figure 6-1A), notwithstanding that most datasets were derived from different samples. We also compared N_d with H' of taxonomic profiles (MetaPhlAn), and observed a similarly high correlation (Figure 6-1B). This indicates that Nonpareil can rank communities by alpha-diversity, in addition to providing accurate projections of coverage.

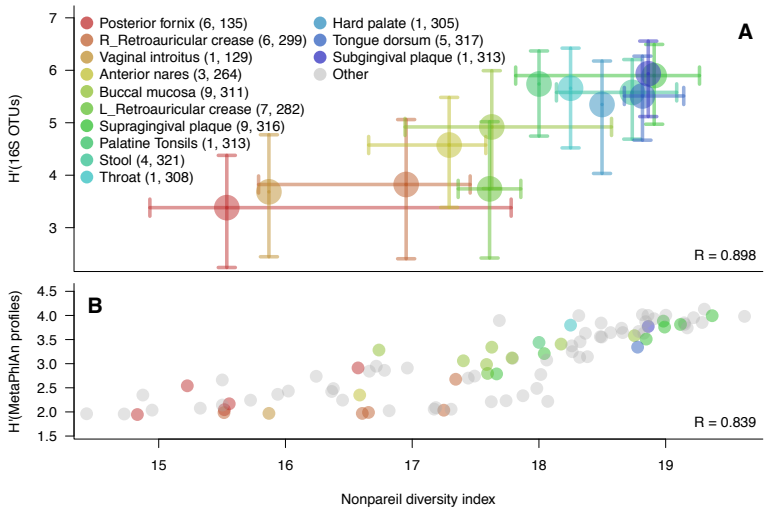


Figure 6-1. Nonpareil's N_d diversity values for 250 HMP datasets.

A. Nonpareil's N_d of metagenomic datasets and Shannon's H' of 16S OTUs for 250 HMP datasets grouped by body site (circles: median, whiskers: 90% interval, number of metagenomic and 16S datasets in parenthesis). **B.** N_d and H' of MetaPhlAn taxonomic profiles.

Availability

Nonpareil 3 is available for online analyses at http://enve-omics.ce.gatech.edu/nonpareil. The code is distributed under the artistic license 2.0 and is available at https://github.com/lmrodriguezr/nonpareil.

Funding

This work has been supported by the U.S. Department of Energy (Award DE-SC0006662) and by the U.S. National Science Foundation (Award 1356288).

Supplementary data

Supplementary data are available in APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 6.

CHAPTER 7: MICROBIAL COMMUNITY SUCCESSIONAL PATTERNS IN BEACH SANDS IMPACTED BY THE DEEPWATER HORIZON OIL SPILL

Originally published on September 2015 in *ISME Journal* 9 (9): 1928-1940, DOI: 10.1038/ismej.2015.5.

Luis M. Rodriguez-R, Will A Overholt[1], Christopher Hagan[2], Markus Huettel[2], Joel E. Kostka[1,3] & Konstantinos T. Konstantinidis.

Although petroleum hydrocarbons discharged from the Deepwater Horizon (DWH) blowout were shown to have a pronounced impact on indigenous microbial communities in the Gulf of Mexico, effects on nearshore or coastal ecosystems remain understudied. This study investigated the successional patterns of functional and taxonomic diversity for over 1 year after the DWH oil was deposited on Pensacola Beach sands (FL, USA), using metagenomic and 16S rRNA gene amplicon techniques. *Gamma-* and *Alphaproteobacteria* were enriched in oiled sediments, in corroboration of previous studies. In contrast to

---

[1] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA.

[2] Department of Earth and Atmospheric Sciences, Florida State University, Tallahassee, FL, USA.

[3] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

previous studies, we observed an increase in the functional diversity of the community in response to oil contamination and a functional transition from generalist populations within 4 months after oil came ashore to specialists a year later, when oil was undetectable. At the latter time point, a typical beach community had reestablished that showed little to no evidence of oil hydrocarbon degradation potential, was enriched in archaeal taxa known to be sensitive to xenobiotics, but differed significantly from the community before the oil spill. Further, a clear succession pattern was observed, where early responders to oil contamination, likely degrading aliphatic hydrocarbons, were replaced after 3 months by populations capable of aromatic hydrocarbon decomposition. Collectively, our results advance the understanding of how natural benthic microbial communities respond to crude oil perturbation, supporting the specialization-disturbance hypothesis; that is, the expectation that disturbance favors generalists, while providing (microbial) indicator species and genes for the chemical evolution of oil hydrocarbons during degradation and weathering.

Introduction

The oil spill caused by the blowout of the Deepwater Horizon (DWH). Drilling rig in April 2010 constitutes the largest accidental release of oil into the marine environment in recorded history. Oil contamination from the DWH spill had a profound impact on indigenous microbial communities, and all available studies recognize shifts in the composition of microbial communities in direct contact with oiled seawater and sediments in comparison with pristine environments (Atlas & Hazen, 2011; Joye, Teske, & Kostka, 2014; King, Kostka, Hazen, & Sobecky, 2015; Kostka, Teske, Joye, & Head, 2014). Moreover, consistent patterns were observed in microbial communities exposed to DWH oil in the Gulf of Mexico including an increase in the relative abundance of members of the *Gammaproteobacteria*, a prevalence of known hydrocarbon-degrading populations, and the enriched abundance and expression of genes related to hydrocarbon degradation (Joye et al., 2014; King et al., 2015; Kostka et al.,

2014). These patterns and microbial responses are also in accordance with observations from laboratory studies and previous accidental releases of oil in marine environments (Berthe-Corti & Nachtkamp, 2010; Greer, 2010; Head, Jones, & Röling, 2006; McGenity, Folwell, McKew, & Sanni, 2012; Röling et al., 2002; Yakimov, Timmis, & Golyshin, 2007).

The Unified Area Command estimated that approximately one-half of the ~4.9 million barrels of oil released from the DWH blowout reached the ocean surface (Lubchenco et al., 2010), and a portion of this surfaced oil transported to nearshore and coastal ecosystems was buried in the sediments (Hayworth, Clement, & Valentine, 2011; P. Wang & Roberts, 2013), impacting approximately 850 km of beaches from east Texas to west Florida (Michel et al., 2013). Oil started depositing on the Pensacola Beach sands studied here on 22 June 2010. The input of large amounts of crude oil, including an array of potentially toxic compounds, posed a potential disturbance for benthic microbial communities (Valentine et al., 2012). Available studies to date were primarily focused on the water column and/or deep sea ecosystems, and less is known about the response or adaptation of sedimentary communities to oiling (Huettel, Berg, & Kostka, 2014). Studies characterizing the taxonomic shifts between contaminated and non-contaminated beach sediments recognized that the oil input strongly affected the beach sand microbial communities, which responded with increased bacterial cell densities (Kostka et al., 2011), reduced taxonomic diversity, and a succession of microbial populations that paralleled the changes in abundance and composition of deposited hydrocarbons (Bik, Halanych, Sharma, & Thomas, 2012; Kostka et al., 2011; Lamendella et al., 2014). Consistent responses have been observed across study sites, although other factors such as site heterogeneity and seasonal fluctuations in environmental parameters have been shown to somewhat confound assessments of the oil impact in certain beaches (Newton et al., 2013), sometimes making them undetectable (Röling et al., 2004). In general, an initial increase in the relative representation of known oil

degraders, mostly of the *Gammaproteobacteria* class (most notably *Alcanivorax*), was observed together with a temporal succession characterized by an increase in relative abundance of *Bacillus*, *Microbacterium* and members of the *Alphaproteobacteria* class at later stages, when recalcitrant oil hydrocarbons predominate (Kostka et al., 2011). Moreover, the increase in oil degraders was concomitant with an increased expression of polycyclic aromatic hydrocarbons, *n*-alkane and toluene degradation genes as assessed by metatranscriptomics (Lamendella et al., 2014). Although these findings provided important insights into the effects of oil on benthic microbial community composition, the gene functions selected for and the genomic adaptations in response to the presence of oil remained mostly uncharacterized in the Gulf coast.

Previously identified shifts in microbial communities in response to DWH oil, both in the water column and sediments, indicated significant susceptibility of these communities; susceptibility defined as the degree to which community composition changes in response to disturbance (Shade et al., 2012). These observations are in accordance with the majority of ecological studies addressing the effect of disturbances such as carbon inputs on microbial communities, which have found evidence of susceptibility (reviewed by (Allison & Martiny, 2008)). However, the magnitude, stability and stochasticity of functional responses, as well as the mechanisms driving the taxonomic and functional composition of the microbial community after disturbance are not well understood (Reed & Martiny, 2007). For example, it has been recognized in plant and animal communities that generalist populations better withstand disturbances, whereas specialist populations tend to be favored in stable environments (specialization-disturbance hypothesis; (Vázquez & Simberloff, 2002)). According to the disturbance-specialization hypothesis, most specialist taxa are selected against when communities experience a severe disturbance, as they are adapted to relatively narrow niches in their natural ecosystem. In contrast, generalists are more

resilient to disturbances altering the niches. In turn, the taxonomic diversity of the community is negatively impacted by a disturbance, but the functional diversity can increase as an effect of the disturbance. Although some previous studies applied ecological theory to describe the response and recovery of community dynamics to disturbance (*cf.* (Prosser et al., 2007; Shade et al., 2012)), the relationship of disturbance and specialization remains largely unexplored in microbial communities. Disturbed communities are typically observed to encompass reduced taxonomic and/or phylogenetic diversity compared with undisturbed controls, but whether this pattern translates to reduced functional diversity or increased specialization remains largely unknown. In this study, we aimed to characterize the response of sedimentary microbial communities from Pensacola Beach to the DWH oil spill, as an *in-situ* experiment of the effects of disturbance on functional and taxonomic diversity.

Materials and Methods

Beach sands were collected at Pensacola Municipal Beach, FL, USA (30°19.57 N, 087°10.47 W) on 6, 10, 20 and 24 May 2010 (before arrival of the oil plume to the shoreline; hereafter, termed pre-oil communities/samples), 30 July 2010 (one month after the oil reached the beach; oiled), 20 October 2010 (when oil constituents were still present in the sand; weathered oiled), and 14 June 2011 (when oil was not visually detectable; recovered; Table 7-1 and Supporting methods). Samples were collected from aerobic beach sediments (oxygen concentrations between 210 and 230 $\mu mol\, l^{-1}$ down to 55 cm depth, which represents >50% of air saturation level) above groundwater level.

Table 7-1. Samples used in this study.

[a] Reads after quality trimming with maximum probability of error of 1% and minimum length of 50 bp, and removal of contamination with adaptor sequences. [b] Samples with oiled and weathered oil status were distinguished based on visual assessment of oil presence. Recovered status was defined based on undetectable levels of hydrocarbons at depths similar to (previously) oiled samples. [c] Sediment temperature between 0 and 50 cm depth presented as mean±one standard deviation. [d] Data for May 2010 not available, presented values were measured in June 2010. *Cf.* temperatures in May 2011: 25.21±2.07.

Designation	**Reads after trimming[a]**	**Status[b]**	**Depth (cm)**	**Sampling date**	**Sediment temp. (°C)[c]**
S1	2 937 972	Pre-oil	0	6 May 2010	29.96 ± 2.66[d]
S2	7 951 456	Pre-oil	0	10 May 2010	
S3	7 837 390	Pre-oil	0	20 May 2010	
S4	6 710 972	Pre-oil	0	24 May 2010	
A	32 840 836	Oiled	30–40	30 Jul 2010	30.49 ± 2.72
B	32 392 430	Oiled	35		
C	25 024 134	Oiled	30–40		
D	21 469 632	Weathered oil	48–62		
E	26 279 070	Oiled	40–45	20 Oct 2010	23.73 ± 2.95
F	34 830 190	Oiled	25–47		
G	39 208 672	Oiled	24–36		
H	25 224 316	Weathered oil	50–55		
I600	33 188 686	Recovered	30–40	14 Jun	31.02 ± 2.79

I606	31 477 910	Recovered	30–40	2011	
J598	31 724 116	Recovered	50–65		
J604	28 119 496	Recovered	50–65		

16S rRNA gene amplicons were sequenced, and the resulting sequences were analyzed as described recently (Poretsky, Rodriguez-R, Luo, Tsementzi, & Konstantinidis, 2014). Trimmed sequences were clustered into operational taxonomic units (OTUs) at 97% similarity using UCLUST (Edgar, 2010), OTUs that represented < 0.005% of the total sequences were discarded (Bokulich et al., 2013) and representative sequences of each OTU were classified using the RDP Classifier at 50% confidence (Q. Wang, Garrity, Tiedje, & Cole, 2007). Shotgun community DNA was sequenced, and the resulting metagenomic reads were quality checked, assembled and annotated as described in the Supplementary Online Material. The level of coverage of the community achieved by each metagenomic dataset was estimated and projected using Nonpareil with default parameters (Rodriguez-R & Konstantinidis, 2014c). Assembled contigs were taxonomically annotated using MyTaxa (Luo et al., 2014). 18S rRNA gene-encoding reads were identified by Metaxa (Bengtsson et al., 2011) with e-value < 0.1 and taxonomically annotated using pplacer and taxtastic (Matsen et al., 2010). Read mapping to estimate the relative abundance of genes and taxa was performed using BLAT with default parameters (W. J. Kent, 2002), considering only the best match with alignment length ≥ 80 bp and identity ≥ 97%. Annotation terms and taxa with significantly different abundance between groups of samples were identified using the negative binomial test as implemented in DESeq2 (Anders & Huber, 2010).

To measure the average number of genes per cell with a given functional annotation (genome equivalents), a set of universally conserved single-copy genes were identified among the assembled gene sequences from the metagenomes. All genes were compared against a collection of 101 HMMs

(Dupont et al., 2012), using HMMER3 (http://hmmer.janelia.org/) with default settings and trusted cutoff, excluding genes for which more than one model represented the same gene family. The median sequencing depth (in reads/bp) of the remaining 91 models was used as the normalizing factor for each dataset. The sequencing depth of genes with a given annotation (see below) was estimated for each dataset (in reads/bp), added up and divided by the normalizing factor of the corresponding dataset.

To identify genes related to oil degradation, gene-specific databases were compiled and manually curated. Sequences for AlkB (alkane hydroxylase) and CYP153 (cytochrome P450 family) were derived from the annotated datasets by (L. Wang, Wang, Lai, & Shao, 2010); sequences for NahA (naphthalene 1,2-dioxygenase) were derived from the set compiled by (Lu et al., 2012); and sequences for ArhA (polycyclic aromatic hydrocarbon dioxygenase) and BBS (benzylsuccinyl-CoA dehydrogenase) were derived from UniRef50 clusters (Suzek, Huang, McGarvey, Mazumder, & Wu, 2007). Putative proteins of the assembled metagenomes were functionally identified using blastp (Altschul et al., 1990) against each reference dataset, with a 250 bit-score threshold. The resulting dataset for AlkB was aligned using Muscle v3.8.31 with default parameters (Edgar, 2004), and the gene phylogeny was reconstructed using RAxML v7.7.2 with GTR model (proteins), gamma parameter optimization, and '-f a' algorithm (Stamatakis, 2006). Putative coding fragments predicted with FragGeneScan (Rho, Tang, & Ye, 2010) on sequence reads were subsequently placed onto the reconstructed tree based on a sequence-to-profile alignment built with Clustal Omega v1.1.0 (Sievers et al., 2011), using the evolutionary placement algorithm (Berger et al., 2011). The same placement strategy was independently applied to the partial sequences of AlkB reported in the study by Smith *et al*. (2013) (GenBank entries KF613175-KF613575).

Diversity was calculated as the true diversity of order one (^1D; equivalent to the exponential of Shannon index). The α and γ components were estimated from

the abundance of categories in a sample and in all samples, respectively, and adjusted for unobserved fractions using the Chao-Shen correction (Chao & Shen, 2003) as implemented in the R package entropy (Hausser & Strimer, 2013). Richness was estimated using the Chao1 index (Chao, 1984), and evenness was calculated as the corrected true diversity of order one (number of equivalent groups) divided by the estimated richness (number of groups).

Crude oil hydrocarbons in the sediment samples were identified by gas chromatography-mass spectrometry using an Agilent 7890A Series GC (Santa Clara, CA, USA), coupled to an Agilent 7000 triple quadrupole MS system, as described previously (Zuijdgeest & Huettel, 2012). The Supplementary Online Material provides further information about procedures and analytical techniques.

All sequencing datasets were deposited in the NCBI Sequence Read Archive under project PRJNA260285 and additional material is available at http://enve-omics.ce.gatech.edu/data/oilspill.

Results

Description of samples and their metagenomes

Concentrations of total petroleum hydrocarbons quantified by gas chromatography-mass spectrometry and visible oil stains monotonically decreased between sampling dates (P-values ≤ 0.05, one-sided t-test; Figure 7-1A). Specifically, the depth-integrated sedimentary inventories of small molecular weight aliphatic and aromatic compounds decreased rapidly from 6 and 1 $mg\,kg^{-1}$, respectively, in July to less than 0.5 $mg\,kg^{-1}$ in October. In contrast, gas chromatography-mass spectrometry profiles revealed that sedimentary inventories of aromatic compounds greater than C8 remained unchanged during this same time frame, whereas aliphatic compounds greater than C6 displayed only a marginal reduction.

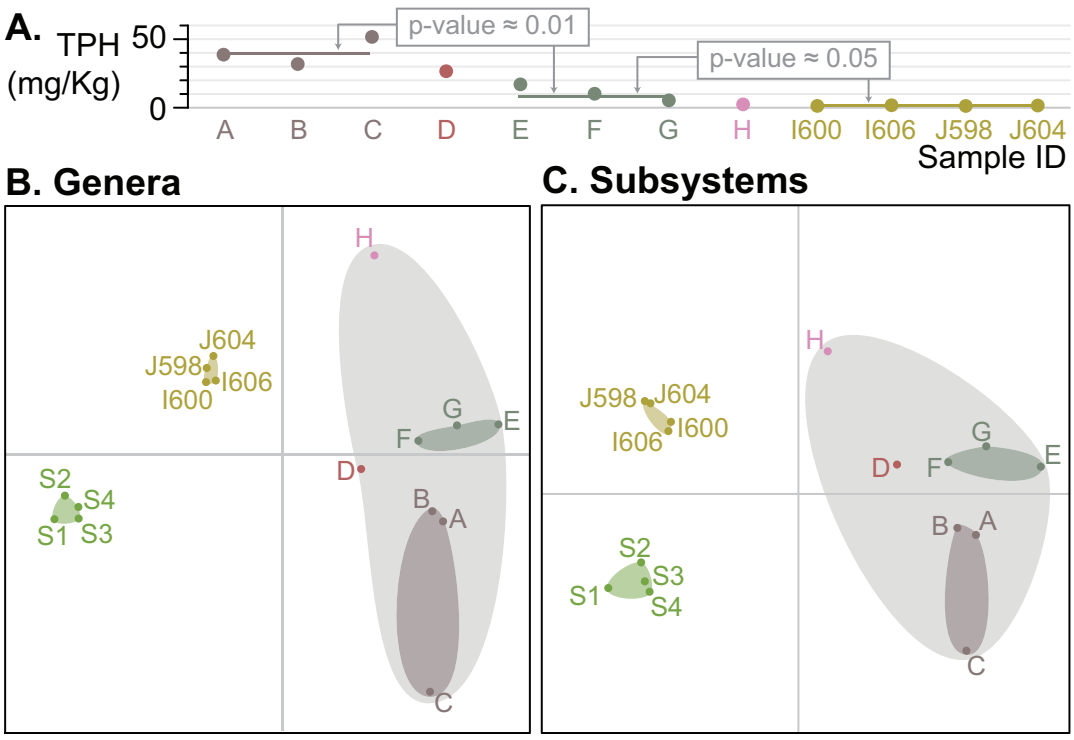


Figure 7-1. Shifts in taxonomic and functional profiles in relation to oil concentration.

(A) The concentration of total petroleum hydrocarbons was significantly higher in July samples (A, B, C) relative to October 2010 (E, F, G), and June 2011 samples (I600, I606, J598, J604). The comparisons between groups (July to October 2010, and October 2010 to June 2011) were performed using one-sided *t*-tests (*P*-values in grey boxes), and the average per group is indicated as horizontal lines. The non-metric multidimensional scaling of **(B)** genera and **(C)** subsystems reveals non-overlapping regions between pre-oiled (green), oiled (shadowed grey) and recovered (olive) samples. The two-dimensional stresses for genera and subsystems are 3.361% and 3.358%, respectively, and the origins are indicated with grey lines. Distance matrices were generated using Bray-Curtis dissimilarities of normalized read counts and ordination was selected by minimizing stress on two dimensions. Note that the heavily oiled samples also form non-overlapping areas by sampling date (dark green and brown), and are distinguishable from weathered oil samples (pink and dark red).

A total of 16 metagenomic samples, ranging in size from 3 to 78 million reads after trimming (paired-end reads with average length of 90–190 bp per dataset), were recovered from each of the four sampling time points, with at least three replicates per time point (Table 7-1). The metagenomes from pre-oil samples had an estimated abundance-weighted average coverage (Rodriguez-R & Konstantinidis, 2014c) of 18–39%, the oiled samples a coverage of 35–60% and the samples from recovered communities an average coverage of 20–25%. Nonpareil curves indicated that the communities in the recovered samples had a higher sequence complexity than both pre-oil and oiled communities, with pre-oiled communities displaying a slightly lower sequence complexity (Supplementary Figure S1A). The described trend in sequence complexity corresponded to the estimated richness of these communities based on OTUs from 16S rRNA gene amplicon data (Supplementary Figure S1B). In general, all metagenomes showed lower sequence complexity than previously determined metagenomes from clayey or silty soils such as rain forest and permafrost but were more complex than freshwater or ocean planktonic metagenomes (Supplementary Figure S1A; *cf.* (Rodriguez-R & Konstantinidis, 2014b)). The July and October 2010 samples (oiled and weathered oil) were assembled into ~56 000 contigs per sample with N50 of ~1400 bp; while those from recovered samples resulted in ~12 000 contigs per sample with N50 of ~745 bp (Supplementary File S1). These results further supported the Nonpareil estimates of higher sequence complexity in the latter samples. In total, ~6 70 000 contigs were obtained with an overall N50 of 1101 bp (723 Mbp in total, from 37 Gbp of sequencing reads), on which ~1.2 million genes were predicted, resulting in an average coding density of 87% (Supplementary File S1).

Microbial community specialization in response to oiling

To assess the temporal effects of the oil spill on the microbial community composition and its recovery, the functional and taxonomic profiles at different time points were compared. Four main groups were identified which significantly

differed in both taxonomic and functional distributions (*P-values* ≤ 0.003, ANOSIM based on Bray-Curtis dissimilarity; Figure 7-1 B and C) and were consistent with the oil concentrations measured *in-situ*: Pre-oil (S1, S2, S3, and S4), Oiled July 2010 (A, B, C), Oiled October 2010 (E, F, G) and Recovered (I600, I606, J598, J604). 16S rRNA gene amplicon data also demonstrated that sample depth played a limited role in structuring microbial communities (Supplementary Figure S2; ADONIS: 3% variance explained by depth *vs*. 75% explained by oiling status and collection date), which was consistent with the facts that the beach sands studied here are subjected to high levels of erosion, and high levels of oxygen (>50% of air saturation level) were detectable at all sampling depths. Hence, our pre-oiled datasets, even though originated from different depths (surficial) compared to oiled datasets (30–65 cm), represented reliable controls for assessing the oiled and recovered microbial communities.

Most notably, the communities exhibited an increase in the functional diversity in oiled samples with respect to pre-oil samples, and a reduction in functional diversity in recovered samples with respect to oiled samples (Supplementary Figure S3A), revealing a different state of lower functional diversity in the recovered communities (Supplementary Figure S3B; DECORANA analysis). Interestingly, this pattern was not observed in the taxonomic diversity, richness or evenness levels (Supplementary Figure S3C-E), indicating that it was primarily due to a decrease in functional specialization of the communities in the oiled samples. This interpretation is further supported by a concomitant decrease in the estimated minimum doubling time in the oiled communities (Supplementary Figure S4A), as expected for bacteria with more generalist strategies (Dethlefsen & Schmidt, 2007). More generalist prokaryotes tend to have larger genomes (Konstantinidis & Tiedje, 2005a), but no significant changes in the estimated average genome size were detected (Supplementary Figure S4B). Nevertheless, these results suggested that the oil disturbance caused community shifts

characterized by a decrease in functional specialization and a consequent increase in functional diversity, which were reversed in the post-disturbance recovery process as the succession advanced.

Oil degradation and toxicity drives community phylogenetic composition

Differences in the composition of the communities from pre-oil, oiled and recovered sediments were detected at various levels of taxonomic resolution (Figure 7-2; Supplementary File S3). At the most general level (domain), recovered communities exhibited higher fractions of eukaryotic and archaeal members than oiled and pre-oiled communities (Figure 7-3A), although no differences in the taxonomic composition of the eukaryotic fraction were observed (Supplementary File S3). The higher fraction of eukaryotic sequences is also consistent with the lower coding potential of May, July and October 2010 metagenomes (~89% of total sequence length was protein-coding) vs the Recovered (June 2011) metagenomes (70%; Supplementary File S1). The higher representation of dominant taxa and lower evenness in communities from oiled samples was also evident at the class level, where *Gamma*- and *Alphaproteobacteria* increased in abundance, with a concomitant decrease of novel taxa (represented by the unclassified fraction; Figure 7-2B). The genera significantly more abundant in oiled than in pre-oiled and/or recovered samples were primarily well-known and suspected hydrocarbon degraders, including *Alcanivorax*, *Pseudomonas*, *Hyphomonas*, *Parvibaculum*, *Marinobacter* and *Micavibrio* (Figure 7-2C). In contrast, groups significantly enriched in recovered samples included taxa typically found in marine environments and known to be highly susceptible to xenobiotics such as the archaeal genera *Nitrosopumilus* and *Cenarchaeum*.

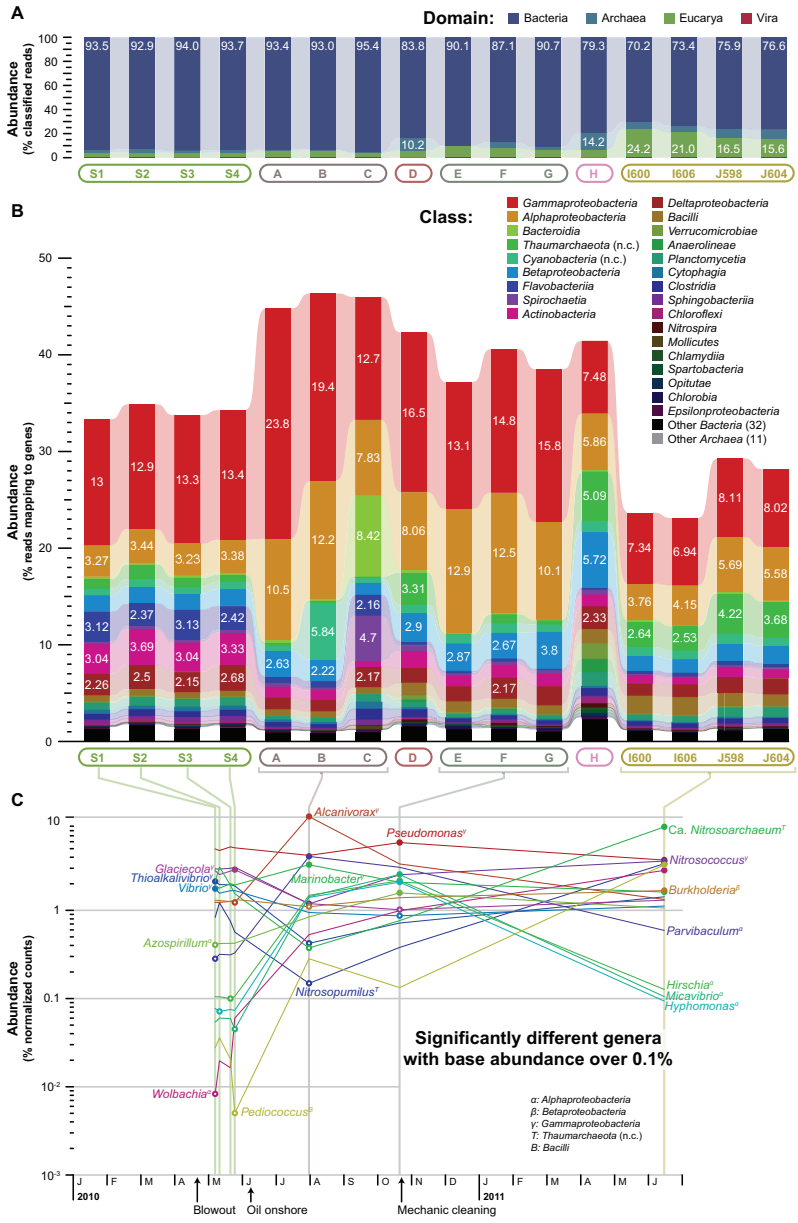


Figure 7-2. Taxonomic shifts in the microbial community in response to oil.

The distribution of metagenomic reads in **(A)** domains and **(B)** classes is displayed for taxa that recruited more than 10% and 2% of the total reads, respectively (white numbers). **(C)** Genera with abundance above 0.1% and significantly different between pre-spill and oiled or between oiled and recovered samples (*P*-value adjusted ≤ 0.01) are also displayed. The minimum and maximum abundance of each genus is indicated with open and filled circles, respectively, and the class is indicated with superscripts.

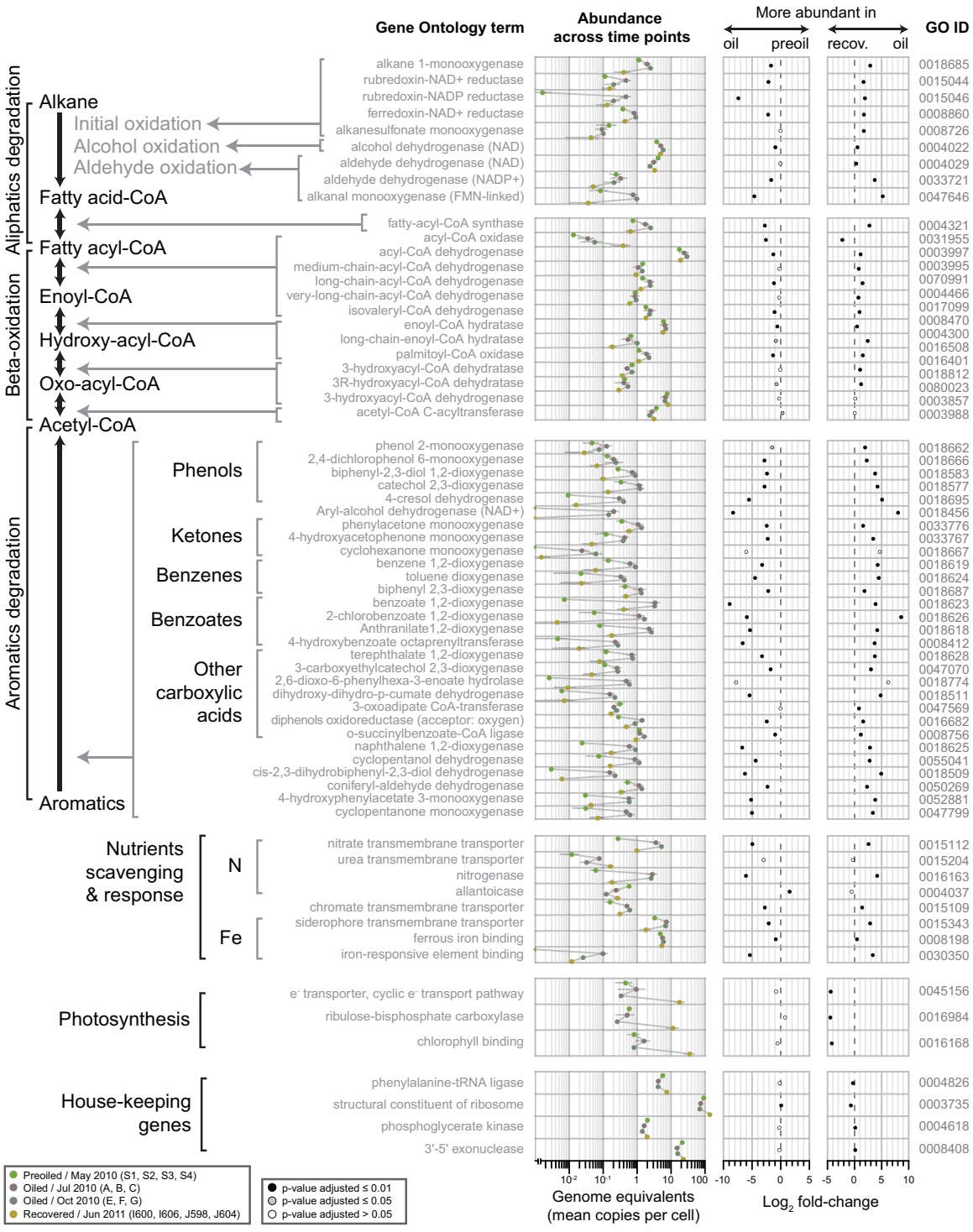


Figure 7-3. Microbial community functional shifts in response to oil.

Selected molecular functions related to hydrocarbon degradation, nutrient scavenging and response, photosynthesis, and some house-keeping genes are listed (left) along with the mean genome equivalents per group of samples (middle) and the $\log_2$ of Preoil/Oiled and Oiled/Recovered fold changes (right). The rightmost column indicates the GO ID of the terms. The abundance was assessed as average genome equivalents (mean copies per bacterial/archaeal cell) on each sampling time (downwards; see legend). The triangles indicate values below the plotted range. The $\log_2$-fold-change was estimated as the $\log_2$ of the ratio of normalized counts between pre-oiled samples (S1, S2, S3, S4) and oiled samples (A, B, C, E, F, G); and between oiled samples and recovered samples (I600, I606, J598, J604). *P*-values were estimated using a negative binomial test.

Functional gene content shift in response to oil

To further investigate the specific functional traits selected by oil presence and, presumably, accounted for the community compositional shifts observed, the abundances of genes associated with alkane and aromatic degradation pathways were compared between pre-oil, oiled and recovered samples. In all evaluated cases, oiled communities displayed a larger prevalence of gene annotations associated with aromatic and alkane hydrocarbon degradation as well as beta-oxidation than pre-oil and recovered communities (Figure 7-3). Interestingly, the relative abundance of most genes associated with aliphatics degradation dropped from July to October 2010, in particular those associated with rubredoxin-NAD+/NADP reduction and aldehyde oxidation (top panel in Figure 7-3). In contrast, the abundance of genes associated with aromatics degradation was roughly maintained or, in some cases, increased from July to October 2010 (second panel in Figure 7-3). In addition, functions related to nutrient scavenging such as allantoicase and nitrogenase (low nitrogen response), and siderophores (iron chelation) were observed to be increased in oiled samples, whereas functions related to primary production such as iron-

responsive elements (iron-responsive binding), as well as functions related to photosynthesis (possibly transported from neighboring marine communities), were enriched in the recovered communities. Notably, most functional categories exhibiting statistically significant difference in abundance in the oiled communities returned to the pre-oil state in the recovered communities, in some cases exceeding their pre-oil levels (Figure 7-3 and Supplementary File S4).

To explore the phylogenetic diversity of genes related to oil degradation, we selected AlkB (alkane hydrolase) as a marker for alkane degradation and reconstructed a high-quality gene phylogeny based on 66 reference genes (mostly based on (L. Wang et al., 2010)) and 43 genes recovered from metagenomic assemblies. In addition, individual metagenomic reads from all datasets were assigned to the most likely node in the tree to provide a quantitative picture of the shifts of AlkB variants over time (Figure 7-4). This dataset included sequences from 14 different genera in five different classes, hence spanning a large diversity of known alkane degraders. Additionally, this dataset covered the diversity of the partial AlkB sequences reported by Smith *et al*. (2013) for the northern Gulf of Mexico, most of which were assigned to clusters IV (73.5%) and II (20.9%). As expected, very few reads from recovered samples were placed in the tree, and most placed reads were derived from oiled or weathered oil samples. However, an intermediate abundance was detected in pre-oil samples (Figure 7-3, first row: alkane 1-monooxygenase). In fact, only cluster III was undetectable in pre-oil samples, whereas all other clusters followed the general trend observed for the entire gene abundance (*cf.* right panel on Figure 7-4 and first row on Figure 7-3). More importantly, the reads from different oiled and pre-oil samples were distributed across different clades, with larger concentrations in a few clades spanning the entire tree, that is, an uneven but phylogenetically unconstrained distribution. Notably, we identified a cluster formed exclusively by genes from this study (labeled ‘OS-I’ in Figure 7-4) most abundant in the pre-oil samples and negatively impacted by the oil spill.

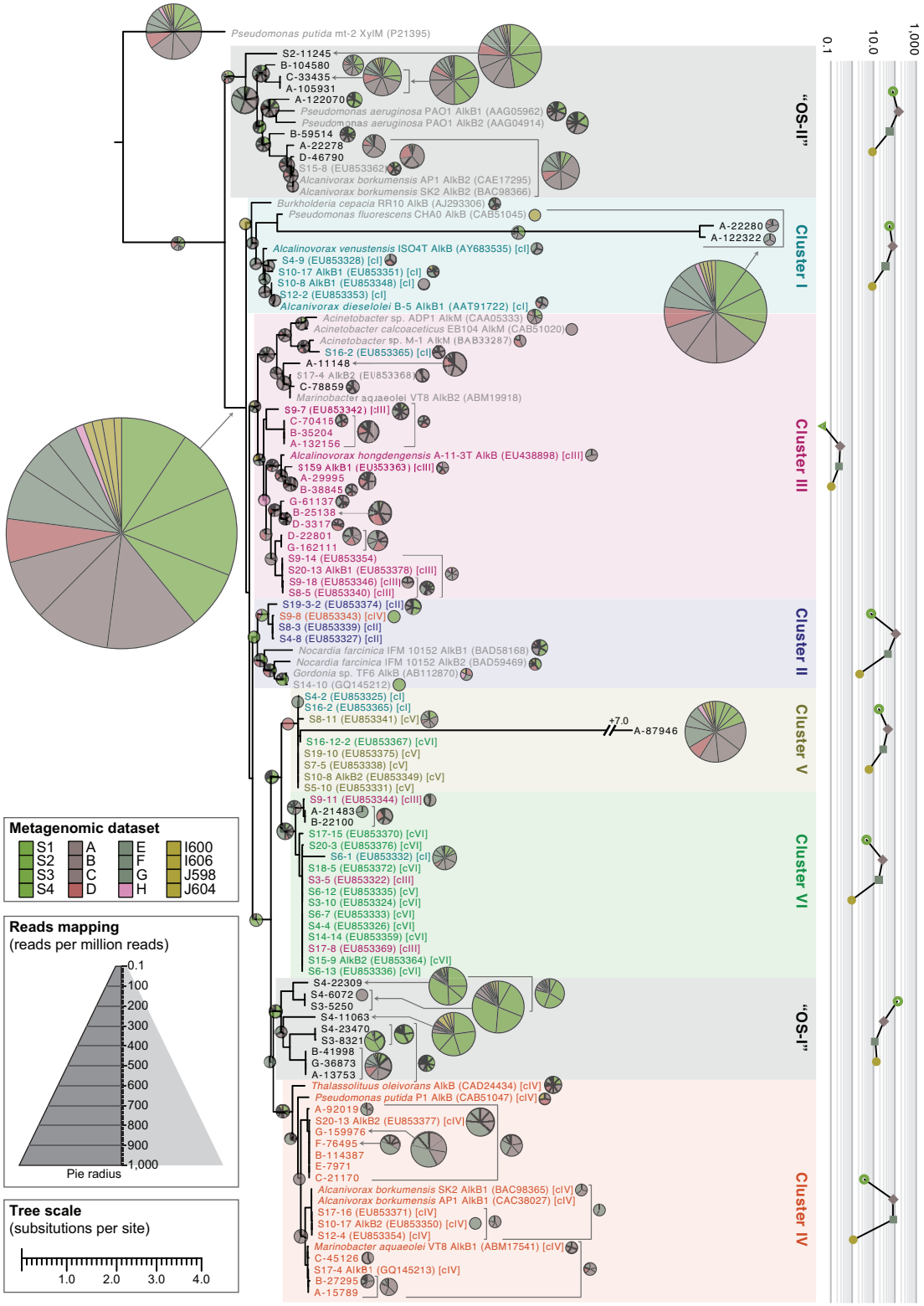


Figure 7-4. Phylogenetic reconstruction of AlkB protein sequences and putative sequences recovered from the metagenomes.

The tree displays reference AlkB (alkane hydrolase) proteins (text colored by clusters, following the nomenclature of (L. Wang et al., 2010)) along with variants assembled from the metagenomes (black text). Proteins with experimental evidence of activity (from heterologous expression or gene knockouts) are indicated by +. Reads mapping to different nodes of the tree are displayed as pie charts. The radius of the pie charts indicates the fraction of the metagenomes mapping to the node (expressed as reads per million, Reads mapping legend), and the different colors of the slices indicate the dataset of origin (Dataset legend). The terminal branch of the sequence A-87946 (cluster V) was shortened by 7.0 units, as indicated by a discontinuity. The right panel indicates the total abundance of each cluster averaged per group of datasets (in reads per million): Pre-oil samples (green), Oiled samples from July 2010 (mauve), Oiled samples from October 2010 (sea green) and recovered samples (olive). Reference sequences (including out-group sequence XylM from *Pseudomonas putida*) and clusters nomenclature (in squared parenthesis) are based on (L. Wang et al., 2010), but the definition of the clusters (colored backgrounds) was broaden to include all sequences in the analysis, and two additional clusters were defined ('OS-I' and 'OS-II').

Population successional patterns and community recovery

In addition to the large differences observed in community composition (both taxonomic and functional) between pre-oil, oiled and recovered samples, the microbial communities characterized in July 2010 differed from those in October 2010 (Figure 7-1B-C), concurrent with a significant reduction in total petroleum hydrocarbons (Figure 7-1A). Examination of the taxonomic distribution revealed that some populations responded rapidly, reaching high abundances in July 2010, with large reductions in abundance by October, and being barely detectable in the recovered samples of June 2011 (Figure 7-3C). These populations included members of the *Alcanivorax*, *Borrelia*, *Spirochaeta*, *Micavibrio* and *Bacteroides* genera. However, some populations were observed to peak in abundance in the October 2010 samples and significantly drop in the

recovered samples, such as *Hyphomonas*, *Treponema*, *Sphingopyxis* and *Hirschia*. Most oil-associated genera did not maintain their abundance in oiled samples of July and October 2010 with the notable exceptions of *Marinobacter* and *Parvibaculum*. The abundance profiles probably reflected organisms with different metabolic properties with respect to oil degradation such as fast responders to easily degradable oil constituents, organisms specialized in degradation of aromatic and more recalcitrant oil fractions, and oil degradation generalists. Finally, we identified a significant increase in minimum doubling time between oiled samples of July and October 2010 based on codon usage bias patterns (Vieira-Silva & Rocha, 2010) (difference of means: 3 h 9 m, *P*-value < 10^{-16}, two-sided *t*-test; Supplementary Figure S4A). The increase in doubling time observed from October 2010 to the recovered samples (June 2011) was much smaller and not statistically significant (difference of means: 27 m, *P*-value: 0.19, two-sided *t*-test). Altogether, these results indicate that the community response to the oil spill involved well-defined successional trends: a rapid response (from May to July 2010), with a peak growth rate in July 2010, followed by a continued decrease in taxonomic diversity (between May and October 2010) and, finally, a reduction in abundance of several known and suspected oil degraders, concomitant with the increase in abundance of several typical marine groups undetectable or rare in oiled samples, a large increase in taxonomic diversity and a decrease in functional diversity.

Discussion

The sands of the Pensacola Municipal Beach received repeated pulses of oil deposition for over a month, and, after about a year, oil was still detected in the beach sands, although it had concentrations below 5 mg kg^{-1} (Figure 7-1A). This indicates that the microbial community faced largely a press (long-term) disturbance given the time scale of microbial generation cycles and migration processes (Shade et al., 2012). Press disturbances often result in community shifts driven by the response traits of individual populations to the disturbance,

presumably sensitivity to toxic compounds and hydrocarbon degradation capabilities in the case of oil contamination. The diversity and abundance of indigenous alkane-degraders preceding the oil spill in the beach ecosystem, as well as the origin of the degraders observed after the spill, was not robustly assessed in previous studies mostly owing to the incomplete diversity recovered in cultures of alkane-degraders and lack of complete understanding of their ecophysiology. The observation of a large and phylogenetically unconstrained diversity of *alkB* genes in the oiled and pre-oil samples supports the hypothesis that the response to the oil spill was primarily caused by shifts in abundances of pre-existing populations, as previously observed in the deep-sea oil plume (Hazen et al., 2010). In other words, the *alkB* genes present in the oiled communities were not derived from a single or a few recent gene alleles but, instead, a large diversity of degraders was latent in the sand and/or surface waters seeping into the sands before the oil spill.

Initial responders (July 2010) included members of the genera *Alcanivorax*, *Borrelia*, *Spirochaeta*, *Micavibrio* and *Bacteroides*, all members of the abundant fraction (≥ 1% of the total community) in the oiled samples. *Alcanivorax* is a genus known for its hydrocarbonoclastic capabilities that can utilize alkanes but not aromatic hydrocarbons (Schneiker et al., 2006); the metabolic capabilities of the other genera in oil hydrocarbon degradation remain speculative. Interestingly, we found putative *alkB* genes (alkane hydrolase) in contigs classified as *Alcanivorax*, *Borrelia* and *Bacteroides*, but no evidence of *arhA* (polycyclic aromatic hydrocarbon dioxygenase) in any of these genera, and putative *nahA* genes (naphthalene 1,2-dioxygenase) only in *Alcanivorax*. The former populations were replaced in the abundant fraction in October 2010 by members of the genus *Treponema* and the class *Alphaproteobacteria* (including *Hyphomonas*, *Sphingopyxis* and *Hirschia*), suggesting a successional dynamic as previously observed based on 16S rRNA gene amplicon data (Kostka et al., 2011; Lamendella et al., 2014, p. 2). Members of the *Hyphomonas* genus have

been reported as abundant members in consortia degrading aromatic compounds, which are typically more recalcitrant components of the crude oil than alkanes and hence, more prevalent in later post-spill stages (Maeda, Ito, Iwata, & Omori, 2010; Maeda, Nagashima, Widada, Iwata, & Omori, 2009). Similarly, *Sphingopyxis* is known to have aromatic hydrocarbon degradation capabilities (Kertesz & Kawasaki, 2010) and was previously detected as a dominant group in soil-derived oil-degrading consortia amended with natural organic matter (Hassan, Taqi, Obuekwe, & Al-Saleh, 2011).

Very few microbial groups, including members of the genera *Marinobacter* and *Parvibaculum*, were consistently enriched in the oiled samples with no noticeable change in abundance between July and October 2010. Putative *alkB* and *cyp153* (cytochrome P450 family) genes, associated with alkane degradation, were identified in assembled contigs assigned to both genera, and putative *nahA* genes, associated with aromatic hydrocarbon degradation, were identified in contigs classified as *Marinobacter*. Members of the *Marinobacter* genus are able to degrade a large variety of aliphatic and aromatic hydrocarbons (Gauthier et al., 1992). Similarly, members of the *Parvibaculum* genus exhibit metabolic capabilities for both aliphatic and aromatic degradation (Lai et al., 2011; Schneiker et al., 2006; L. Wang et al., 2010). In contrast to previous analyses based on 18S rRNA gene amplicons (Bik et al., 2012), no consistent, statistically significant shifts in the taxonomic composition of the eukaryotic fraction were detected between sampling dates or degree of oiling (Supplementary File S3).

Finally, in June 2011, *Synechococcus*, *Pediococcus* and archaeal genera including *Nitrosopumilus*, *Cenarchaeum* and *Nitrosoarchaeum* dominated the abundant fraction (in contrast to oiled samples), and an overall increase in the eukaryotic fraction was observed. Many of the former microbial groups are abundant in oligotrophic or nutrient-poor marine ecosystems, indicating that they represent the sensitive fraction of the community to the oil spill, but to a large extent the community was resilient, as generally observed in microbial

communities (Allison & Martiny, 2008). Notably, the observed succession process exhibited signs of community recovery, but the community in June 2011, 1 year after the oil reached the shoreline, significantly differed from that in May 2010, before oiling, similar to the results of previous microcosm experiments on oil amendment of beach sediment inocula (Röling et al., 2002). The differences between the recovered community and its counterpart before the oil spill may be due to the long-term effects of the oil disturbance (for example, establishment of new taxa), stochastic events or other environmental factors such as organic matter input, nutrient input and salinity changes. Clearly, more samples and analyses would be required to obtain further insights into the latter issue. Nonetheless, our results also suggest that these sensitive marine groups could serve as indicator species of oil presence and toxicity in future oil spill studies, and thus, potentially provide useful information for guiding bioremediation efforts and decisions by site managers.

In general, microbial communities changed both taxonomically and functionally after exposure to a range of petroleum hydrocarbon concentrations. The community shifts caused a decrease in taxonomic diversity during May to October 2010, with a significant recovery by June 2011 (Supplementary Figure S3C). Interestingly, the functional diversity was observed to follow a contrasting trend: it increased between May and July 2010, was maintained between July and October 2010, and significantly decreased in June 2011 (Supplementary Figure S3A-B). We hypothesize that several oligotrophic (specialized) taxa were strongly outcompeted upon deposition of oil onshore. Growth arrest due to limited hydrocarbon degradation capabilities and/or sensitivity to toxic compounds from the continued presence of oil onshore would impact more severely oligotrophic and/or specialist than copiotrophic and/or generalist populations. Hence, a significant reduction in taxonomic diversity but not functional diversity was expected, as observed in these communities. Moreover, we provided evidence indicating that specific fast-growing organisms (typically assumed to be

copiotrophic) thrived in the presence of relatively high concentrations of petroleum hydrocarbons. In other words, the disturbance favored generalist organisms in the communities, and the post-disturbance communities were characterized by a narrower set of more specialized functions. This observation seems counterintuitive because a common expectation is that a press disturbance would exclusively select for few highly specialized organisms, in this case oil-degraders. Nevertheless, this trend is predicted by the disturbance-specialization hypothesis (Vázquez & Simberloff, 2002), is consistent with ecological succession theory and was previously observed in plant communities (for example, (Bazzaz & Pickett, 1980)). It should be mentioned, however, that the patterns observed here might be specific to disturbances from crude oil and sand beach ecosystems and not immediately generalizable to other, more selective disturbances (Röling & van Bodegom, 2014). Crude oil is composed of tens of thousands of different carbon sources that would favor generalists in early succession, as well as toxic compounds that would preferentially select against specialists. Sand beaches in the Gulf of Mexico are characterized by low carbon content and nutrient-poor conditions relative to marshes or other coastal sediments (Huettel et al., 2014). This could explain the relative abundance of putative chemolithoautotrophic archaea (*Nitrosopumilus*, *Cenarchaeum*) in the recovered communities and suggests a suppression of a range of organisms that are adapted to carbon-limited conditions.

In summary, the community response was primarily characterized by two concomitant trends. First, most of the community is selected based on the ability to survive under disturbed conditions, that is, the response to the disturbance correlates negatively with the level of specialization. Second, few organisms with traits selected by the disturbance become highly abundant, as niche opportunity promotes invasion (Pintor, Brown, & Vincent, 2011; Shea & Chesson, 2002). Overall, our results provide evidence of complex successional patterns in the studied communities, involving invasion promoted by capabilities for oil

hydrocarbon degradation, as well as population survival generally hindered by specialization and susceptibility to oil toxicity, and a general recovery of diversity, specialization and sensitive marine groups a year after the disturbance.

Supplementary data

Supplementary data are available at http://www.nature.com/ismej/journal/vaop/ncurrent/suppinfo/ismej20155s1.html.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation (NSF) award no. 1241046 (to KTK), OCE-1057417 and OCE-1044939 (to MH, JEK), the NSF graduate research fellowship no. 2013172310 (to WAO), and by a grant from BP/The Gulf of Mexico Research Initiative to the Deep-C Consortium (#SA 12-12, GoMRI-008). We thank Patrick Chain and the personnel of the Los Alamos National Laboratory for sequencing of the samples.

CHAPTER 8: BIOGEOGRAPHY AND SEASONAL VARIATION DISENTANGLED IN MICROBIAL META-COMMUNITIES OF FIVE CONNECTED LAKES

Reproduced with permission from **Luis M. Rodriguez-R, Despina Tsementzi**[1]**, Chengwei Luo**[1]**, Janet K. Hatt**[1] **& Konstantinos T. Konstantinidis**. All copyright interests will be exclusively transferred to the publisher upon submission.

Abstract

The importance of biogeography and seasonality for community assembly has long been recognized in aquatic microbial communities. However, their relative effect on biogeography and the distinction between historical and contemporary effects remain open questions. Here, we present a framework in which we explicitly consider the effects of multiple factors on α- and β-diversity using database-independent techniques like read redundancy and *k*-mer compositions to derive the γ-diversity components of a meta-community, and then evaluate the variance between sampling sites explained by different independent variables. Using this framework, we analyzed a collection of 70 Illumina metagenomic datasets from five lakes and two estuaries along the Chattahoochee River

---

[1] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

(Southeast USA) between 2010-2015, including bimonthly samples from the most-upstream lake, Lake Lanier (GA), and annual late-summer samples of all sites. We evaluated the effects of biogeography, seasonality, annual variation, and primary production (Chlorophyll A), among other factors, on community diversity. Variations in α-diversity were largely explained by seasonality (54%) and geographic distance (40%), significantly increasing downstream and during the winter. Similarly, observed β-diversity was explained mainly by seasonality (49%) and geographic distance (39%) with modest contribution from the landscape composition (9%), and revealed at least three lacustrine microbial provinces. Endemic genera to these provinces (or habitat specialists) included members of *Epsilonproteobacteria* and *Bacilli* endemic of Lake Lanier, and members of *Gammaproteobacteria* associated to estuarine locations. Taxa with strong seasonal rhythm (with annual periodicity) included members of *Cyanobacteria* associated with late summer. Our framework allowed us to explicitly quantify and disentangle the seasonal and biogeographic effects on microbial meta-community diversity of this interconnected freshwater ecosystem.

Introduction

Recognizing the ecologic processes and factors that govern community assembly at different scales is a central goal in ecology. Recent advances in molecular techniques have allowed the exploration of microbial community structure, paving the way to characterize such processes in the richest and most abundant clades on Earth (Handelsman et al., 2007; Zinder & Salyers, 2015). Moreover, parallel developments in ecologic theory have enabled the quantitative evaluation of ecologic processes in extant communities (Hubbell, 2008; Weiher et al., 2011). The application of these innovations to the study of microbial community assembly has created intense debate, emerging from controversial conclusions related to -for instance- the relative effect of biogeography and stochastic processes on community assembly or the link between geographic distance and structure in soil microbial communities (Martiny et al., 2006; Ofiţeru

et al., 2010; Rousk et al., 2010; Xiong et al., 2012). Studies of microbial community diversity patterns can reveal not only the existence of biogeographic structure but also the underlying processes shaping microbial communities. For example, a classic debate in microbial community assembly relates to the existence of a global seed bank under contemporary environmental selection (Baas Becking, 1934). This debate is importantly informed by the conceptualization of different microbial habitats, regions with distinct environmental characteristics reflecting on communities through contemporary processes, and microbial provinces, geographic units with significant dispersal restrictions setting communities apart by historical processes (Martiny et al., 2006). The distinction between microbial provinces and habitats is non-trivial in practice and may be confounded by the fact that relatively isolated locations that could form provinces tend to have characteristic environmental properties forming unique habitats. However, since the earliest assessments of geographic distance impact on genetic distance (Cho & Tiedje, 2000), mounting evidence indicates that biogeographic patterns similar to those previously described for marcoorganisms are exhibited by microbial communities both with respect to environmental factors as well as geographic distance, although the distinction is seldom explicitly made (Fierer & Lennon, 2011; Green & Bohannan, 2006; Hanson, Fuhrman, Horner-Devine, & Martiny, 2012; Lindström & Langenheder, 2012; Martiny et al., 2006; Ramette & Tiedje, 2006). This suggests that distinct microbial provinces may indeed exist, although contradicting evidence of a global-scale microbial seed banks has also been found (Gibbons et al., 2013). Interestingly, a recent evaluation of global-scale bacterial assemblages demonstrates that the most abundant taxa from diverse habitats are typically cosmopolitan (Nemergut et al., 2011). Consequently, the identification of microbial provinces, even when clearly delimited, could be dampened by community assessment techniques with low genetic resolution or high abundance detection limits. On one hand, most evaluations of community differences to date have been based on sequences of the small subunit of

ribosomal RNA genes (16S rRNA) (Martiny et al., 2006). For example, 16S rRNA has been used to explicitly identify bacterial provinces at a global scale (Nemergut et al., 2011). Indeed, 16S rRNA sequences are tractable and well suited for the global description of communities, but filters operating in species sorting from regional to local scales may operate near or below the species level, a degree of resolution that cannot be fully resolved with 16S rRNA alone (Konstantinidis & Tiedje, 2005b). For example, recent evidence from population studies indicates that taxonomic resolutions attainable with fragments of 16S rRNA genes might be insufficient to detect finer-scale processes such as adaptive radiation (Bahl et al., 2011; Chrismas, Anesio, & Sánchez-Baracaldo, 2015). On the other hand, the community coverage of molecular datasets is seldom evaluated leaving the question of detection limits an uncertainty (Rodriguez-R & Konstantinidis, 2014b). Moreover, variations in between samples are often evaluated with coverage-dependent methods, and commonplace procedures to alleviate this coverage effect often cause decreased statistical power (McMurdie & Holmes, 2014), while different data treatment and transformations have a significant impact on the measurement of the effect different conditions have on community assembly (Yannarell & Triplett, 2005).

Despite these limitations, the strong effect of seasonality has been previously demonstrated in aquatic communities, both in freshwater and marine environments (Caporaso, Paszkiewicz, Field, Knight, & Gilbert, 2012; Fuhrman et al., 2006; Fuhrman, Cram, & Needham, 2015; Gilbert et al., 2012; Giovannoni & Vergin, 2012; Hanson et al., 2012; Poretsky et al., 2014). However, the evidence on the effect of historical factors like dispersal is less clear. A previous multi-scale study in European lakes found that the effect of contemporary local factors drive freshwater community assembly, with little or no influence of dispersal limitations or historic factors detected even at continental scales (Gucht et al., 2007). However, this study was conducted in communities characterized by fingerprinting through denaturing gradient gel electrophoresis (DDGE), a method

with limited resolution at the species and sub-species levels. Similarly, variations in bacterial assemblages from temperate humic lakes detected by automated ribosomal intergenic spacer analysis (ARISA) were found to be driven by intrinsic phytoplankton dynamics, which in turn was found to respond mainly to environmental conditions, but no explicit evaluation of geographic distances was performed (A. D. Kent, Yannarell, Rusak, Triplett, & McMahon, 2007). In fact, the evaluation of dispersal limitations or its effects on community diversity in freshwater habitats is still poorly understood (Brendan Logue & Lindström, 2008), although there is strong evidence showing that dispersion plays an important role in community assembly at regional scales for select bacterial groups (Barreto, Conrad, Klose, Claus, & Enrich-Prast, 2014; Glaeser & Overmann, 2004; Oda, Star, Huisman, Gottschal, & Forney, 2003). Moreover, variations in freshwater communities have been observed to be stronger between different water bodies than within the same lake, with intermediate variations between interannual samples of the same lake, and minimal variation between sampling locations of a lake (Jones, Cadkin, Newton, & McMahon, 2012; Newton, Jones, Helmus, & McMahon, 2007). More explicitly, unmeasured geographically-structured environmental parameters and/or limited dispersal were found to strongly influence the taxonomic turnover in an unconfined aquifer (Stegen et al., 2013). We expect such inter-site migrations to be dominated by water streams, since it has been observed that only moderate rates of immigration operate from air to freshwater ecosystems with nearly undetectable influence in community dynamics (Jones & McMahon, 2009), while migration through streams has been found to be nearly unconstrained (Fierer, Morse, Berthrong, Bernhardt, & Jackson, 2007).

Here, we use metagenomic datasets to attain high-resolution dissimilarities between communities in a meta-community of five interconnected freshwater lakes and two estuarine locations along the Chattahoochee River in the Southeastern USA between 2010 and 2015. The use of metagenomic datasets in

the evaluation of α- and β-diversity has been hindered by technical limitations owing to the difficulty of building accurate taxonomic profiles from whole-genome sequences using non-comprehensive databases. Here, we bypass this limitation by leveraging direct sequence read comparisons previously demonstrated to attain species-level resolution (Ondov et al., 2016; Rodriguez-R et al., Under review). We explicitly model the effect of seasonality, geographic distances, environmental conditions, and incomplete sampling in this chronoseries with spatial resolution, demonstrating the rhythmic patterns and the estiblishment of provinces in this aquatic meta-community. We also discuss seasonal preferences and endemism of particular taxonomic groups.

Methods

Geographic Characterization

The basin of the Chattahoochee/Flint/Apalachicola River (henceforth simply Chattahoochee River) was determined based on the elevation-derived hydrologic data of World Hydro Reference Overlay (Esri Hydrology Team, 2012) at scale 1:288,895 using ArcGIS, and the characteristic basins irrigating each lake were determined with the same overlay at scale 1:144,488 (Lehner, Verdin, & Jarvis, 2008). On each basin, land cover was derived from MDA USA Information Systems (Esri Inc., 2015) at scale 1:288,895 (http://arcg.is/1PTOfVU). The types of land use areas quantified by this platform include **Deciduous and Evergreen forest**: trees with >3 m height; **Shrub/Scrub**: Woody vegetation <3 m in height; **Grassland**: Herbaceous grasses; **Agriculture**: cultivated crop lands; **Water**: All water bodies greater than 0.08 ha; and **High and medium Density Urban**: Areas with over 70%, or between 30 and 70% of constructed materials that are a minimum of 60 m wide (asphalt, concrete, buildings, etc.). The land cover areas were measured using ImageJ 1.49v color thresholding combined with particles analysis (Schneider, Rasband, & Eliceiri, 2012) with images at 15.75 pixels per km (~64km/pixel). For each lake, characteristic basin cover was measured as the

cover profile in the basin irrigating the lake minus the basin irrigating the lake immediately upstream; *i.e.*, each region of the map is only measured once for the lake immediately downstream. Cumulative basin cover was defined as the cover profile in all the basin irrigating the lake; *i.e.*, the sum of the cover profiles for the target lake and all other upstream lakes.

Sampling and Metadata Collection

All samples were collected from the lower epilimnion (typically 3-5m depth) of Lakes Lanier (GA), West Point (GA/AL), Harding (GA/AL), Eufaula (GA/AL), and Seminole (GA/FL) at least 10 m away from the littoral zone, and two locations in the Apalachicola estuary off the coasts of Apalachicola and East Point (FL). Each sample was accompanied by *in situ* measurement of physicochemical parameters using a portable water quality meter (Global Water) including water temperature, pH, oxidation/reduction potential, turbidity, dissolved oxygen, and salinity, as well as *ex situ* spectroscopic measurement of chlorophyll A. Additional nutrient concentration measurements were taken *in situ* or *ex situ* using a portable spectrophotometer (Hach DR/2010) and various Test 'N Tube Nutrient sets (Hach) following manufacturer recommendations for selected nutrients, including sulfate, nitrate, nitrite, ammonia, sulfide, phosphorous, chlorine, chemical oxygen demand, and total organic carbon. Samples consisted of 20 l water collections using a horizontal sampler (Wildco Instruments).

DNA Extraction and Sequencing

Water samples were immediately stored at 4°C and processed typically within 1-4 h, and no more than a day post collection. Water was sequentially filtered with a peristaltic pump through 2.5 μm and 1.6 μm GF/A filters (Whatman), to capture large particles and eukaryotic cells, and cells were eventually captured on 0.2 μm Sterivex filters (Millipore). Thus, all sequenced metagenomes represent the 1.6-0.2 μm cell size fraction, with the exception of samples LLGFA_1308A and

LLGFA_1309A, which represent the 2.5-1.6 μm fraction. Filters were preserved at -80°C. DNA extraction was performed as previously described (DeLong et al., 2006) with minor modifications. Briefly, frozen filters were fragmented and placed in microcentrifuge tubes with lysis buffer (50 mM Tris-HCl, 40 mM EDTA, and 0.75 M sucrose) and 1 mg/ml lysozyme, and incubated at 37°C for 30 min. Reactions were subsequently incubated with 1% SDS, 10 mg/ml proteinase K, and 150 µg/ml RNase for 4 h at 55°C in a rotating hybridization oven. DNA was extracted from lysate with phenol and chloroform, precipitated with ethanol, and eluted in Tris-EDTA (TE) buffer. On average, DNA yield was 1.7 µg per liter of water filtered. One nanogram of DNA was used to prepare sequencing libraries using the Nextera XT Kit described by the manufacturer (Illumina). Libraries (9-11 pM) were sequenced using Illumina GA II, HiSeq, or MiSeq following manufacturer's protocols.

Quality Control of Metagenomic Datasets

All sequenced metagenomic datasets were subjected to quality control and those not passing minimum requirements were re-sequenced. Sequencing reads were trimmed using SolexaQA++ (Cox et al., 2010) with minimum PHRED quality score of 20 and minimum fragment length of 50 bp, and clipped to remove residual sequencing adaptor contamination (if any) using Scythe (https://github.com/vsbuffalo/scythe). Abundance-weighted average coverage of the datasets was estimated using Nonpareil with default parameters (Rodriguez-R & Konstantinidis, 2014c). A minimum size of 1Gbp after trimming and 50% coverage were required for all datasets in this study.

α- and β-diversity Estimation

Diversity in the metagenomic datasets was estimated at the read level to avoid database biases using tools that have demonstrated relative robustness to sequencing effort. The N_d index of α-diversity was estimated using Nonpareil with

default parameters and alignment kernel (Rodriguez-R et al., Under review). Briefly, Nonpareil estimates the abundance-weighted average coverage per sequencing effort in metagenomic datasets and determines the sequence diversity as the inflexion point of the curve in logarithmic scale. Dissimilarity between datasets was estimated as Mash distances (Ondov et al., 2016) with sketch size 10,000 and *k*-mer size 21. Briefly, Mash distances are estimated using MinHash, an algorithm for the heuristic comparison of *k*-mer profiles using the first occurrences in the profile (sketch) sorted by deterministic hashing functions. The Mash distance is defined as a transformation of Jaccard distances between sketches. Additionally, we estimated the minimum expected distance between replicates by simulating an idealized random sampling scheme as follows. ***(i)*** A community with abundance profiles (in numbers of cells) following a log-normal distribution (Preston, 1948) was estimated with Curtis' technique (Curtis, Sloan, & Scannell, 2002) with N_T/N_{max} = 6.0 based on previous characterizations of Lake Lanier communities (Oh et al., 2011; Poretsky et al., 2014), a total of 1.5×10^{21} cells (based on cell counts indicating a concentration of about 9×10^5 cells in 750 μl of water and a total lake volume of 1.3 km^3), and a minimum species abundance of 1 cell. ***(ii)*** The sub-sampling fraction was estimated taking into account the effects of water sampling, DNA extraction, library preparation, sequencing effort, and trimming/clipping of sequences. Multiplying these effects, we estimate that each sample represents on average about 5×10^{-20} the total community DNA, and we use this as sampling probability to randomly generate two samples from the community in *i* using binomial sampling or Poisson or Normal approximation whenever possible. ***(iii)*** The two sampled profiles are compared using the Jaccard index and the value is transformed to reflect Mash units (Ondov et al., 2016). ***(iv)*** The entire procedure *i-iii* is repeated 1,000 times and the average resulting distance is used as an estimation of minimum distance between replicates; *i.e.*, the expected difference between independent samples derived from the exact same community. This procedure was implemented in a Ruby library available at

https://github.com/lmrodriguezr/sampleton. Note that the sample distances obtained with this transformation are an overestimation (or rather an upper-boundary estimation) of the expected Mash distances due to possible overlaps in *k*-mer distributions between closely related (but distinct) species. Moreover, the actual makeup of simulated samples assumes random unbiased sampling with respect to taxa, which (if untrue) would further increase simulated distances with respect to those from real samples.

Quantification of Biogeography and Seasonality Effects

The biogeographic patterns were studied with the definition of two variables: fluvial distance to Lake Lanier (the most upstream sampled location) and type of ecosystem (lacustrine or estuarine). All geographic distances between datasets were measured as fluvial distances (*i.e.*, travel distances along the river) in km. The seasonality of the diversity was studied with two techniques. First, a graphic approach was developed, in which the diversity was plotted as a function of the sampling dates expressed in radians (with 1 year corresponding to a full circumference). The complete dataset was bootstrapped 1,000 times, cubic splines were calculated for each replicate, and the resulting average and 90% confidence intervals were reported. Second, for a quantitative analysis, the seasonality was decomposed in two variables: vernality (sine of date in radians) and wintriness (cosine of date in radians), and the effect of these variables, year, and the absolute date in radians on sequence diversity were evaluated on α- and β-diversity (see below). For completeness, the variables autumny and aestivality were also defined simply as the negative of vernality and wintriness, respectively. Distances were summarized by nonmetric multidimensional scaling (NMDS) in two and three dimensions, and metadata was cross-correlated with ordination scores for graphical representation (bi-plots). The effect of temporal, spatial, and physicochemical variables on α-diversity was evaluated using a general linear model coupled with analysis of variance (GLM-ANOVA) for N_d sequence diversity. The effect of the same variables on β-diversity was evaluated using a

distance-based redundancy analysis coupled with analysis of variance (dbRDA-ANOVA) with ten thousand permutations based on the dissimilarity matrix of Mash distances. The above analyses were performed using the implementations available in the R package vegan v2.4-1 (Dixon, 2003).

In addition, Mash distances were modeled using a general linear model (GLM) fit by least squares method. We constructed a first general model where the contribution of habitat (*H:* estuarine *vs*. lacustrine habitat identity indicator function, with values 0 or 1), geographic distance (*Δg*: fluvial travel distance in thousands of km, with range 0-0.8), angular date difference (*Δs*: $[1-\cos(2\pi\Delta t)]/2$, with range 0-1), and sampling date difference (*Δt*: difference of sampling times in years, with range 0-5.3) on Mash distances (with range 0-1) were estimated by their coefficients. A second model specific for Lake Lanier was constructed. This model didn't have any geographic component (habitat or distance), hence including only the temporal variables (*Δs* and *Δt*).

Taxonomic profiles and identification of periodicity and endemism

Metagenomic datasets were independently assembled using IDBA-UD with default parameters (Peng, Leung, Yiu, & Chin, 2012) and scaffolds of at least 500bp in length from all datasets were combined in a single long-contig collection. Co-assembly and re-assembly attempts (even between most similar datasets) didn't yield noticeable improvements. Coding sequences were predicted using MetaGeneMark.hmm with default parameters (Zhu, Lomsadze, & Borodovsky, 2010). Predicted proteins were used to determine the most likely taxonomic affiliation of each contig using MyTaxa with 0.5 score threshold (Luo et al., 2014) in combination with DIAMOND in blastp mode with minimum score 60 and a maximum of 5 target sequences (Buchfink, Xie, & Huson, 2015). Sequencing reads from each metagenomic dataset were mapped against the combined long-contig collection using BLAT in fastMap mode and only best-matches were considered (W. J. Kent, 2002). For each dataset, the number of

reads mapping to contigs of any given genus in the database were estimated using MyTaxa.distribution.pl (Luo et al., 2014). Format manipulations were performed with the utilities of the enveomics collection (Luis M Rodriguez-R & Konstantinos T Konstantinidis, 2016).

In order to identify periodic taxa, abundance of genera per sample (only in Lake Lanier samples) was estimated as sequencing read counts normalized by cumulative sums transformation (Paulson, Stine, Bravo, & Pop, 2013). Seasonality per genus was estimated by generating cubic smoothing splines of log-normalized abundance by sampling date (ignoring year). In order to implement the cyclic nature of the data into the spline analysis, all the abundances were used in triplicate with sampling dates in tandem; *i.e.*, with sampling times in radians in the ranges [-2π , 0], [0, 2π], and [2π, 4π]. Only the central fragment of the splines was used in the range [0, 2π]. Genera with Pearson's correlation index between spline-derived abundance and observed abundance below 0.5 (p-value 0.0016) were discarded as non-periodic. For the periodic genera, normalized abundance of 24 points uniformly distributed across the year was estimated from the splines (*i.e.*, approximately biweekly points) and clustered using Ward's hierarchical clustering method on correlation distances ([1-R]/2).

The same methods described above were applied to the metagenomic samples from the different lakes from the months of August, September, and October. Cubic smoothing splines were estimated on log-normalized abundance per fluvial distance from Lake Lanier, and genera with Pearson's correlation index between spline-derived abundances at lake locations and observed abundances below 0.5 (p-value: 0.00034) were discarded as non-endemic, *i.e.*, either uniformly rare or cosmopolitan across sampled locations.

Results

Chattahoochee River Basin Land Cover

The sites in this study are located along the Chattahoochee River in the Southeastern USA (Figure 8-1A). The drainage basin of the river (Figure 8-1B) extends from the north of Georgia, along the border with Alabama, and meets the Gulf of Mexico in northern Florida (main riverbed highlighted in grey). The basin is mainly covered by deciduous and evergreen forests, with about 10-20% coverage from urban areas (Figure 8-1C). Agriculture covers a significant portion of the characteristic basin of Lake Seminole (*i.e.*, the fraction of the basin between Lake Eufaula and Lake Seminole), and to a lesser extent a portion of Lake Eufaula, but is rare in other segments of the valley. Wetlands and Woody wetlands contribute a significant fraction of the basin South of Lake Seminole, and along the shoreline beaches (barren/minimal vegetation) become commonplace.

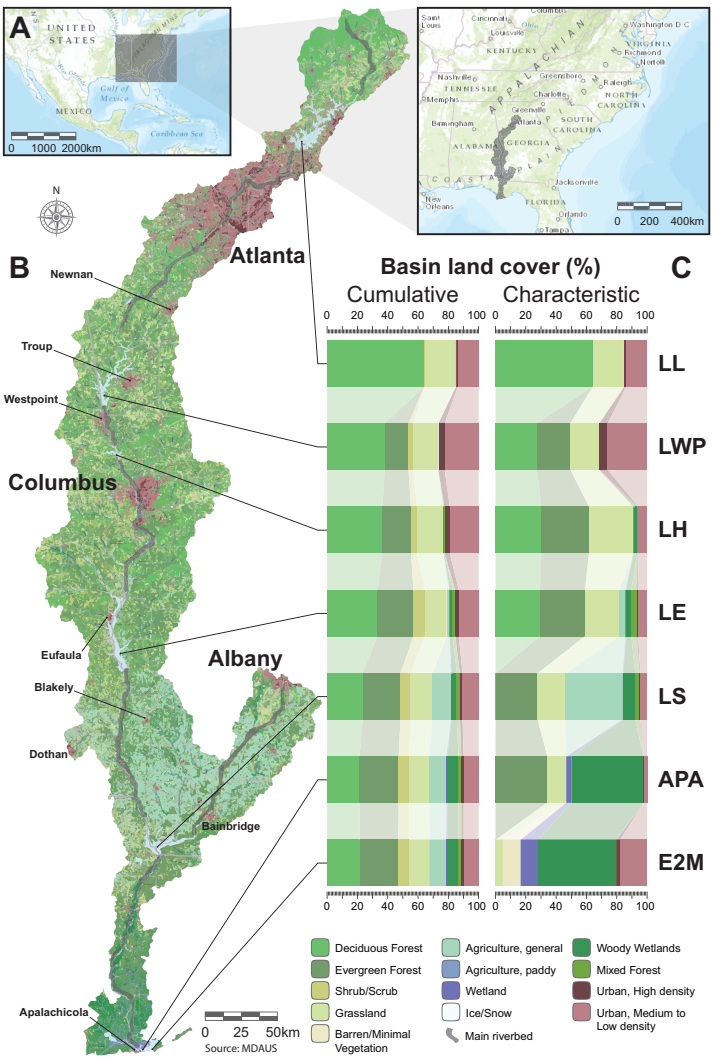


Figure 8-1. Geographic location and land use along the Chattahoochee River basin.

A. Location of the river drainage basin in the Southeast continental USA. **B.** Land use cover map of the river basin. The main riverbed of the Chattahoochee River is shown in the context the river's drainage basin. The sampling locations included all five major lakes and two estuary sites. **C.** Land use profiles for each sampling locations. Land use profiles are shown for each sampling location, estimated based on the upstream drainage basin for each lake (characteristic) or the cumulative land use consisting from all the upstream locations for each site.

Biogeography along the Chattahoochee River

Sampling locations along the Chattahoochee River included five lakes and two estuarine locations spanning 761 km of the riverbed between most-upstream Lake Lanier and the river mouth in the Gulf of Mexico. Yearly sampling in all locations was conducted between 2010 and 2015 (except 2011) between the months of August and October (with additional bimonthly sampling in Lake Lanier). Metagenomic datasets presented a much higher similarity within the same sampling location than between locations, with 72% of variance explained by sampling location (dbRDA-ANOVA, p-value $<10^{-5}$). Moreover, distances between groups per sampling site reflected their geographic locations, with site clusters driven by location along the Chattahoochee River (Figure 8-2A). This pattern was interrupted by estuarine samples, which presented an additional degree of separation from lacustrine samples presumably due to the influence of marine waters (Figure 8-2B). In lacustrine samples, fluvial distance between locations significantly correlated with dissimilarities between datasets (Figure 8-2C), with 39% of the Mash distance variance being explained by geographic distance alone (dbRDA-ANOVA, p-value $<10^{-5}$). We estimate that, on average, lacustrine samples differ by an additive 1% Mash units for every one hundred kilometers distance between them, with additional 10% Mash units between lacustrine and estuarine samples. Interestingly, this pattern in β-diversity was accompanied by a downstream increase in α-diversity (Figure 8-3) with a significant linear accumulation of about 0.07 N_d units every one hundred kilometers (including estuarine samples) resulting in 0.54 additional N_d units on average along the entire riverbed (or 0.40 N_d units between the most distant lakes). This is a modest but highly significant increase (Pearson's R = 0.651, p-value $2x10^{-6}$) with 40% of the variance on N_d sequence diversity explained by distance to Lake Lanier (GLM-ANOVA, p-value 10^{-3}).

In order to distinguish between the effect of geographic distance and habitat variation, we tested the effect of pH, turbidity, temperature, conductivity, dissolved oxygen (DO), total dissolved solids (TDS), chemical oxygen demand (COD), total organic carbon (TOC), and concentrations of chlorophyll A, ammonia, sulfide, and reactive phosphorous on β-diversity. As mentioned above, the difference between lacustrine and estuarine habitats had a strong effect on dataset dissimilarities beyond that of distance alone. However, among lacustrine samples, the habitat variation was relatively minor, and distances alone remained much more informative on the observed β-diversity. After controlling for sampling date (in radians, modulo 1 year), only four physicochemical variables had a significant effect on dataset dissimilarity (dbRDA-ANOVA, p-values < 0.01, variables independently assessed): conductivity, turbidity, DO, and TDS. However, conductivity, turbidity, and TDS had a strong linear correlation with each other and formed clines along the river, and none remained significant after controlling for geographic distance. Notably, the strongest associations between these variables and dataset dissimilarity were found for conductivity and TDS (var. explained 31% in both cases, independently assessed). The latter two variables highly correlated to each other (Pearson's $R = 0.9999$) and with location along the riverbed (Pearson's $R = 0.85$) mainly driven by significantly lower values in Lake Lanier (t-test, p-value 4×10^{-10}). Re-evaluation of the effect of these variables on dataset dissimilarities excluding Lake Lanier rendered all four non-significant even without controlling for geographic distance (dbRDA-ANOVA, p-values > 0.01), although geographic distance itself remained a strong predictor of Mash distances (dbRDA-ANOVA, p-value $<10^{-5}$, var. explained 22%). Finally, in order to identify potential environmental effects due to the surroundings but not reflected by the measured parameters, we tested the influence of the land cover composition of the characteristic basin of each lake on community variation. After controlling for geographic distances in lacustrine datasets, we identified several land cover types with significant influence in Mash distances (dbRDA-ANOVA, p-values < 0.01), including agriculture (general and paddy), wetland, mixed forest,

and urban (high and medium-low density). However, the fraction of the characteristic basin of each lake covered by these types explained only 9.6% of the community variations, compared to 39% of geographic distance in the mixed dbRDA model. Altogether, these results suggest that individual lakes represented distinctive microbial provinces but only two habitats types were distinguishable (lacustrine and estuarine). The distinction between Lake Lanier and all the other lakes could represent a secondary habitat, but no other habitat distinctions were found.

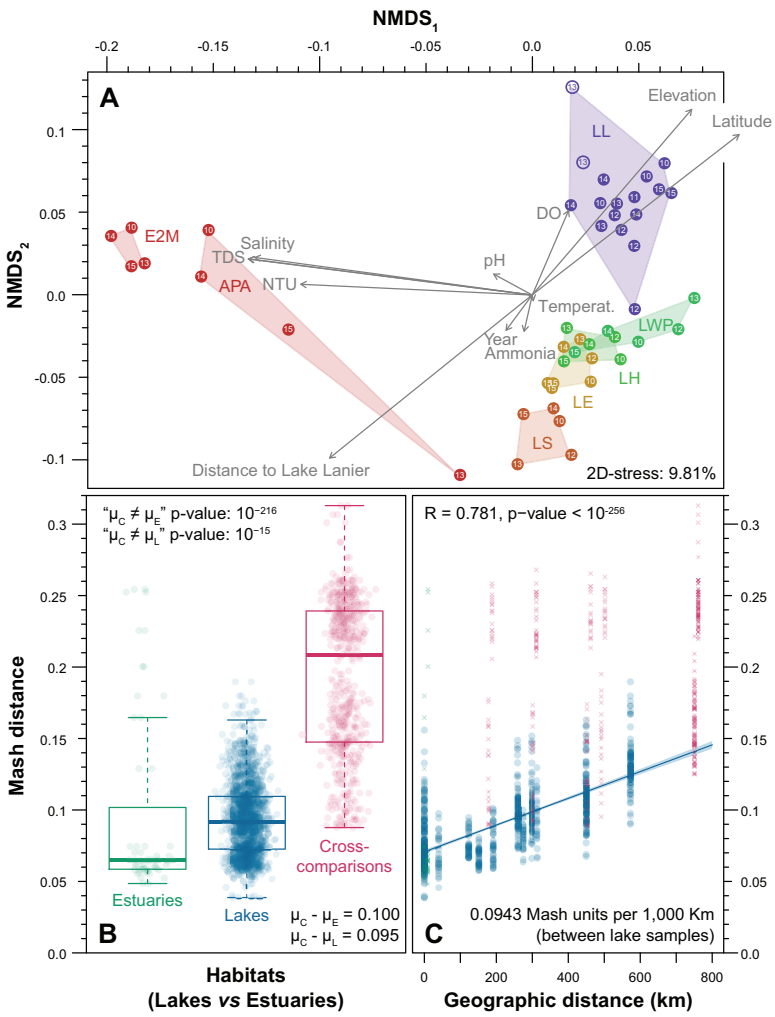


Figure 8-2. Geographic clustering of samples.

Geographic distances (measured along the river) and type of ecosystem significantly impact observed β-diversity. E2M and APA sites are estuarine, and all others are lacustrine. Only August-October datasets were used from Lake Lanier to eliminate the impact of seasonal variations in this analysis. **A.** The 2-dimensional NMDS on the different samples is colored and grouped by site, with year of sampling indicated within each dot. In addition, the correlation of the ordination scores with location, sampling year, and some environmental parameters is displayed as grey vectors. The Chattahoochee River connects all sampling sites and flows southwards; hence all three measurements have high correlations with each other. Datasets derived from larger particle fractions (2.5-1.6 μm) are shown as empty circles. The samples distinctly separate by site, and site clusters deploy by location throughout the river. **B.** Mash

distances between samples discriminated by habitat. Green dots indicate pairs of estuarine samples, dots in blue indicate pairs of lacustrine samples, and pink dots represent pairs of one estuarine and one lacustrine sample. The p-values for the equality of means and the average difference between cross-comparisons and the other sets are also shown. **C.** Correlation between geographic distance and Mash distance between pairs of lacustrine samples (blue dots). The solid blue line and band indicates the average linear fit and 95% confidence interval. The Pearson's R and it's corresponding p-value are also displayed. Pairs of lacustrine samples (green crosses) and cross-comparisons (pink crosses) are shown for completeness.

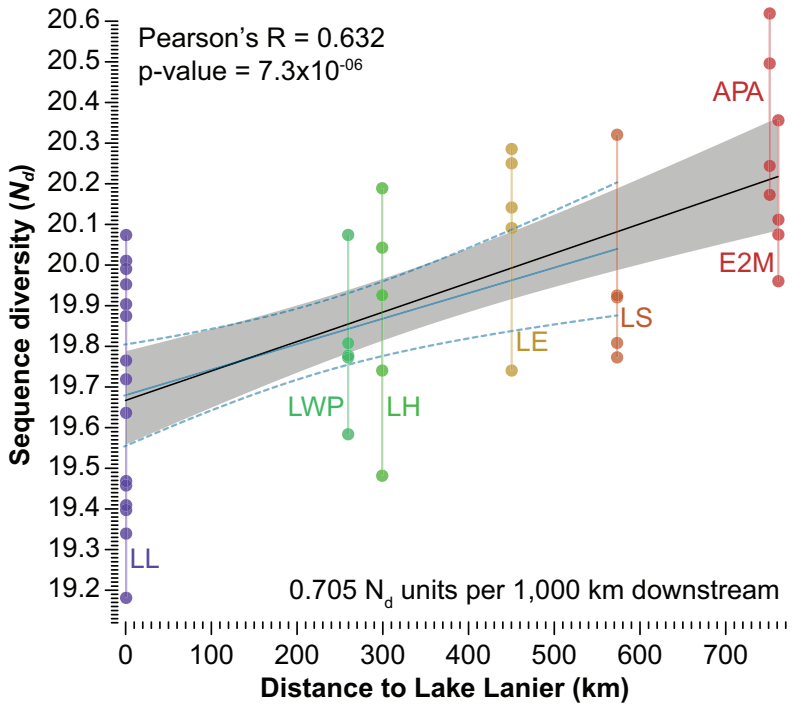


Figure 8-3. Geographic variation of α-diversity.

The sequence diversity (N_d) increases downstream the Chattahoochee River. The distance to Lake Lanier is measured as distance throughout the river stream and significantly correlates with the Nonpareil sequence diversity Index (p-value <0.001). The linear model fit is shown (black solid line) together with the 95% confidence interval (grey band). Only August-October samples from Lake Lanier were included to eliminate the

effect of seasonality on this analysis. The same analysis excluding estuarine samples results in a similar correlation (solid and dashed blue lines). Datasets derived from larger particle fractions (2.5-1.6μm) present higher sequence diversity (Aug/13: 20.269, Sep/13: 20.747) and were excluded from this analysis.

Seasonality of Lake Lanier

Lake Lanier (northernmost sampled location; Figure 8-1) was sampled nearly bimonthly between 2010 and 2015, allowing a detailed characterization of seasonal patterns in the local community. Samples from this location presented a distinctive annual rhythmical pattern (Figure 8-4A), with seasonality decomposition variables alone explaining 49% of the variance on Mash distances (dbRDA-ANOVA; Wintriness/Aestivality: 21% var. explained, p-value $<10^{-5}$; Autumny/Vernality: 28% var. explained, p-value $<10^{-5}$). In addition, small cumulative interannual changes in the communities were detected, with year deploying perpendicular to the seasonality plane in 3-dimensional NMDS (data not shown) and an additional 3.1% variance explained with weak support (dbRDA-ANOVA, p-value 0.056). This periodicity was accompanied by a repeating pattern on α-diversity with datasets exhibiting larger N_d sequence diversity in the winter and minimum values on or between late summer and early fall (Figure 8-4B), resulting in an expected 0.85 N_d units of difference between the dates when diversity is maximum (early January) and minimum (late July). Seasonality decomposed variables explained 54% of the variations in N_d sequence diversity (GLM-ANOVA; Autumny/Vernality: 19% var. explained, p-value $6x10^{-4}$; Wintriness/Aestivality: 35% var. explained, p-value 10^{-5}).

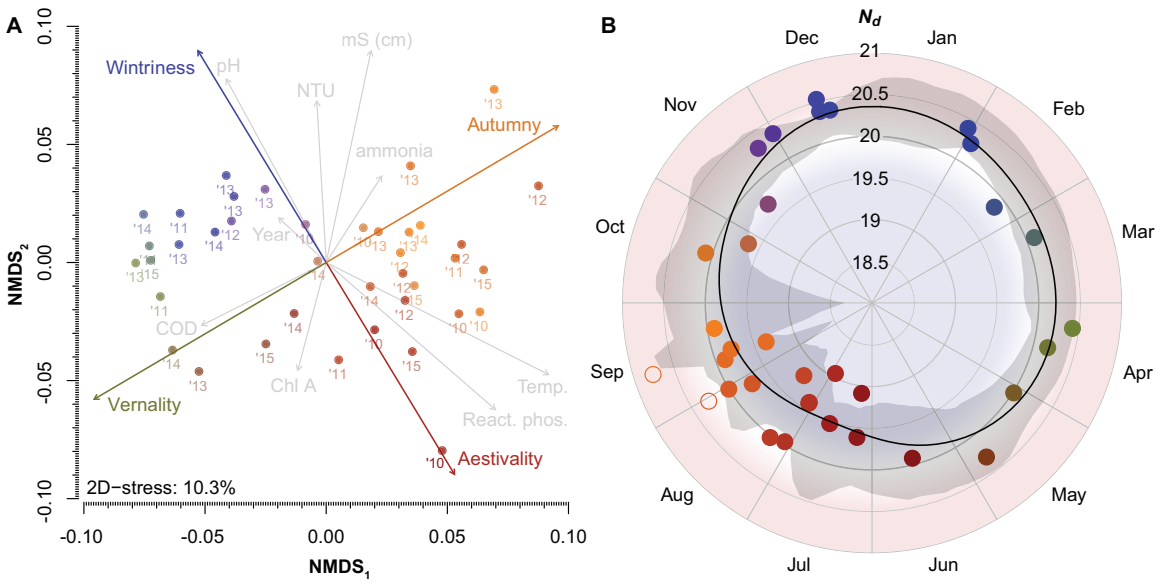


Figure 8-4. Seasonal variation on α- and β-diversity in Lake Lanier.

A. 2-dimensional NMDS ordination with datasets colored by sampling date using color interpolations between blue on January 1st, green on April 2nd, Red on July 2nd, and orange on October 1st. Each dot is labeled with the sampling year. In addition, vectors of metadata are overlaid indicating their correlation with the ordination axes. Vectors of seasonality decomposition are colored by the day of their maximum value (as described above). Note that datasets deploy following a seasonal progression counterclockwise. **B.** Seasonal variation on α-diversity in Lake Lanier. Each sample is located at a rotation determined by sampling date, and length (distance from the center) determined by sequence diversity (N_d). Colors indicate the season as in **Figure 8-4**. Datasets derived from larger fractions (2.5-1.6 μm) are presented for completeness as empty circles, but were excluded from the statistical analyses. All other data points were interpolated using cubic smoothing splines with 1,000 bootstraps. The average (black solid line) and 90% confidence interval (grey band) of N_d are presented. The background indicates the average diversity (19.891) in white, and values above in red and below in blue. The sequence diversity of the datasets presented a seasonal variation, with winter samples (dark blue) exhibiting the largest diversity and summer/fall samples (dark red/orange) the lowest.

Modeling β-diversity

In order to compare the relative contributions of seasonality and biogeography to the differentiation between datasets, we generated a general linear model (GLM) including all the variables detected in the previous sections. The resulting equation allows the prediction of Mash distances between two datasets given their sampling characteristics (Figure 8-5). More importantly, it allows the generalized comparison of relative effects in common units:

$$Mash\ distance_{complete} \approx \frac{7.5 + 7.2H + 6.2\Delta g + 3.4\Delta s + 0.11\Delta t}{100}$$

Equation 8-1

Where *H* is a habitat identity indicator function, *Δg* is the fluvial travel distance (in thousands of km), *Δs = [1-cos(2πΔt)]/2* represents angular date difference, and *Δt* is the difference of sampling times (in years). Interestingly, both geographic distace and seasonal components present similar variation ranges in Mash units (Figure 8-5 A *vs*. B). The geographic distance (*Δg*) has a coefficient of 6.2, resulting in an average prediction range of 0.047 Mash units, while the seasonality alone (*Δs*) has a coefficient of 3.4 corresponding to a range of 0.034 Mash units (or 0.039 when including absolute date difference as in Figure 8-5A; *Δt*). This indicates that both seasonal variations and sample locations contribute similarly to the detected dissimilarities between datasets, while habitat type had a stronger effect with a coefficient of 7.2 corresponding to a range of 0.072 Mash units. The expected distance when all independent variables were zero was 0.075 Mash units, about twice as large as expected by chance between replicates (0.040). The same procedure applied on Lake Lanier samples alone resulted in similar coefficients, with a slightly higher predicted effect of seasonality and smaller expected distances between samples, probably owing to the lower technical variation accumulated in this reduced set:

$$Mash\ distance_{LL} \approx \frac{6.8 + 4.6\Delta s + 0.15\Delta t}{100}$$

Equation 8-2

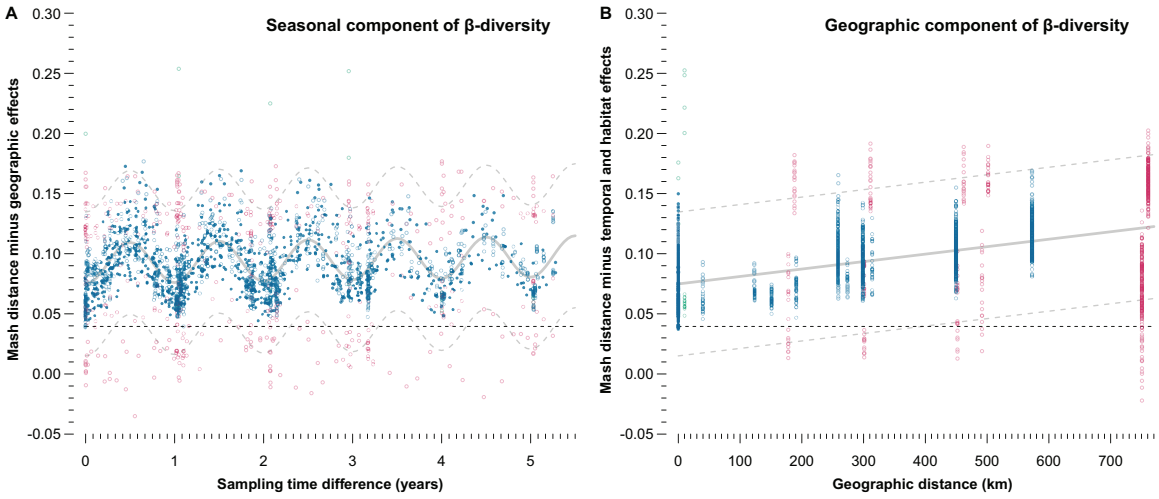


Figure 8-5. β-diversity modeling.

A general linear model including temporal and geographic variables was estimated. Dataset comparisons between different habitats are represented as pink circles, comparisons between lacustrine samples as blue circles, and comparisons between estuarine samples as green circles. Comparisons within Lake Lanier are shown as filled circles, all other comparisons are represented as open circles. The expected distance between replicates is shown as black horizontal dashed lines. **A.** Mash distances between datasets were transformed to remove the estimated effects of geographic distance and habitat difference. The resulting values are presented as a function of sampling time difference. The model predictions are shown assuming no geographic differences (solid gray line) together with 95% prediction intervals (dashed gray lines). **B.** Conversely, Mash distances with removed estimated temporal and habitat effects are presented here as a function of geographic distance. The model predictions and 95% prediction intervals were estimated assuming not temporal or habitat differences and are presented in grey solid and dashed lines, respectively.

Periodicity and endemism in community members

Estimated genus profiles of Lake Lanier samples were used to identify taxa presenting repeatable inter-annual patterns (periodic). From 948 genera detected in Lake Lanier, 78% (746) were selected as periodic by comparing year-round smoothed profiles (annual values superimposed) with observed abundances (Pearson's $R > 0.5$, p-values < 0.0016). The high number of detected periodic taxa indicates that the observed seasonal patterns in the above sections were driven by synchronic changes in the community with a wide taxonomic base, as opposed to large changes in only few abundant taxa. Biweekly estimations of abundances per periodic genus are presented in Figure 8-6A. Four distinct clusters formed, corresponding to genera peaking in early and late summer (spring-summer: n=147, summer-fall: n=249), and early and late winter (fall-winter: n=102, winter-spring: n=248; Figure 8-6A). In order to identify associations at higher taxonomic ranks with particular seasonality groups (including non-periodic), we evaluated the expectation that the number of genera per class (or phylum, when class was unavailable) was simply determined by the fraction of genera in each group using a chi-square test per taxonomic group and corrected the resulting p-values by false discovery rate (Figure 8-6B). Seven taxa had corrected p-values ≤ 0.05, namely the phylum *Cyanobacteria* and the classes *Clostridia*, *Alphaproteobacteria*, *Actinobacteria*, *Bacilli*, *Spirochaetia*, and *Thermomicrobia*. None of the groups appeared to be enriched in the non-periodic collection (Figure 8-6B). Notably, photoautotrophic genera from the phylum *Cyanobacteria* were generally associated with summer months. From 40 genera detected in this phylum, 5 typically peaked in early summer and 26 in late summer.

Similarly, we applied this technique to the distributions of all samples between the months of August through October by geographic distances in order to identify endemism in the different locations (or in sets of nearby locations; Figure 8-6C). From 952 detected genera, 75% (712) were selected as preferentially

associated to particular transects of the river, *i.e.*, displaying endemism. Expectedly, detected endemism was mainly associated with the differentiation between habitats, with 272 genera being most typically associated with estuaries and 440 with all or most lakes (Figure 8-6C). Finally, we detected two additional groups of endemism that appeared to be associated with the upstream Lake Lanier (n=67) and the downstream Lakes Eufaula and Seminole (n=54). The latter group was further clustered into sets of genera typical of Lake Seminole (n=32), Lake Eufaula (n=7), and no evident preference (n=15). Eleven classes were detected with non-uniform genera counts (chi-square; p-values ≤ 0.05; Figure 8-6D). *Cyanobacteria* appeared to be enriched in the Lake Lanier group, but no classes particularly associated to the downstream lakes were detected. Classes mainly associated to the estuarine group included *Alphaproteobacteria*, *Gammaproteobacteria*, *Flavobacteria*, and those more common in the non-endemic group included *Gammaproteobacteria* and *Bacilli*.

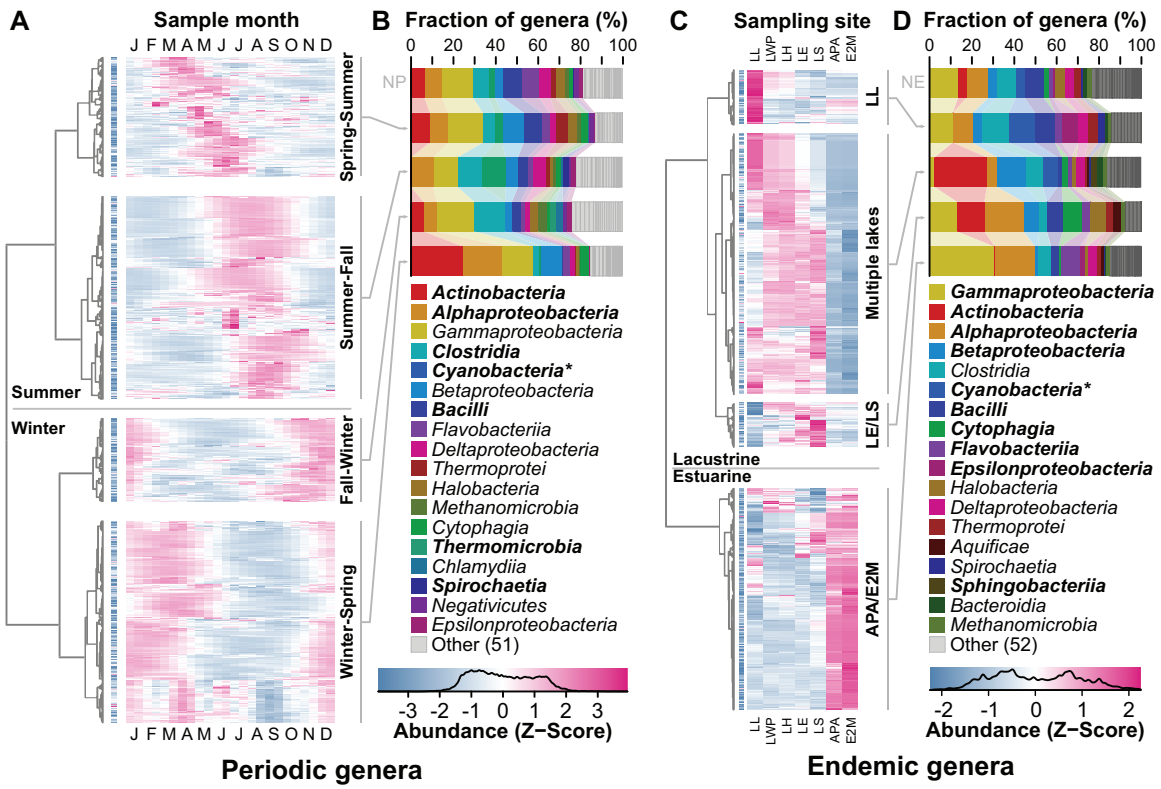


Figure 8-6. Abundance profiles and taxonomic affiliations of periodic genera in Lake Lanier and endemic genera in sampling sites.

A, B. Cubic smoothed splines of normalized log-abundance of genera in Lake Lanier per sampling date (in radians) were used to summarize the rhythmic patterns of periodic taxa (Pearson's R between observed data and cubic splines above 0.5). **A.** The predicted biweekly abundance throughout a model year (columns) of each periodic genus (rows) is displayed as log-abundances in Z-scores per row (see legend). Genera were hierarchically clustered using correlation distances and the ward method, and four top-level groups were identified roughly corresponding to early and late summer (top) and early and late winter (bottom). **C, D.** Similarly, cubic smoothed splines were estimated with respect to fluvial distance to Lake Lanier in August through October samples, and genera displaying endemism were selected (Pearson's R between observed data and cubic splines above 0.5). **C.** The predicted abundance per location (columns) of each endemic genus (rows) is displayed as log-abundances in Z-scores per row (see legend). Hierarchically clustering using correlation distances and the ward method were used to separate four groups, three associated to lakes (top) and one to

estuaries (bottom). **B, D.** The class distribution of the genera on each cluster and in the non-periodic (NP) or non-endemic (NE) groups was estimated when available; the lowest taxonomic rank available above class was used otherwise (e.g., *Cyanobacteria*). Taxa with non-uniform genus counts between groups (Chi-square; p-values ≤ 0.05) are shown in bold typeface.

Discussion

The effect of geographic distances, in contrast to environmental characteristics, on microbial freshwater community assembly is a topic of active debate, while both conceptual and technical limitations may obscure the detection and differentiation of microbial habitats and provinces. On the other hand, it is generally accepted that seasonality plays an important role in structuring of aquatic communities, but quantitative assessments in parallel to biogeographic evaluation are still rare. In this study we analyzed a collection of 69 metagenomic samples from five lacustrine and two estuarine sites connected through the Chattahoochee River, spanning a large region of the Southeastern USA between 2010 and 2015, including 47 samples from the seven different sites collected between the months of August and October, and 39 samples from Lake Lanier collected year-around. In order to leverage the high genetic resolution achievable by whole-genome shotgun approaches while circumventing the limitations of taxonomic approaches caused by incomplete databases, we estimated the α and β components of diversity using recently proposed techniques, based on read-level comparisons (Ondov et al., 2016; Rodriguez-R et al., Under review). The results revealed a strong effect of habitat type in community assembly, with clear differentitation of lacustrine *vs*. estuarine communities and higher diversity harbored by the latter. More interestingly, we observed a significant distance decay between local communities that could not be more satisfactorily explained by any of the measured environmental characteristics, indicating that microbial provinces have formed in this interconnected system by historical (mediated by dispersal) rather than contemporary factors. Moreover, we recognized a

downstream linear increase in community diversity along the riverbed. Although it is possible that the increase in diversity is associated with an unmeasured ecocline, this observation is directly compatible with the interpretation of slow directed migrations, *i.e.*, small dispersal only (or mainly) through the riverflow. Rapid dispersal would uniformize the extant diversity, while undirected migrations (*e.g.*, through air) would break the monotonicity in the pattern. At least three provinces were recognizable among the lacustrine locations, namely an upstream province in Lake Lanier, a mid-basin province in Lakes West Point, Harding, and Eufaula, and a downstream province in Lake Seminole. However, even within the second province, geographic distances had a structuring effect suggesting reduced but detectable historical effects in these three localities. In the upstream and downstream provinces we were able to detect at least 67 and 32 genera displaying significant endemism, respectively. The higher number of genera recognized as endemic of Lake Lanier could partially be the result of a larger set of samples from this location, allowing the assembly of identifiable genomic fragments from rare members, and effectively lowering the detection limit in this location. However, endemic members of each province were likely captured in this dataset and their complete genome sequence, which could provide cues for the underlying factors of endemism, may be recoverable with advances in sequence assembly, including graph representations and/or species binning techniques. Finally, we demonstrated that the effect of geographic distance on freshwater community assembly is lesser but of similar order of magnitude as the widely recognized effect of seasonality. For example, geographic distance was estimated to impact differences between communities with an amplitude of 0.047 Mash units, compared to 0.034 of seasonality. Similarly, geographic distances had an average effect on sequence diversity with an amplitude of 0.54 N_d units, compared to 0.85 by seasonality. Taxonomic identification of drivers of these differences displayed similar results, with 75% of genera identified as endemic (by habitat or province), while 78% of genera identified as displaying rythmic annual patterns of abundance. More

quantitatively, geographic distance explained 39% of variation among samples from August to October as assessed by dbRDA, while 49% of variation among Lake Lanier samples could be explained by seasonality. Together, our results revealed that the studied metacommunity is shaped by the interaction between historic and contemporary effects largely mediated by restrictions in dispersal and seasonal environmental variations, with modest contribution of landscape composition and only minor contributions of environmental characteristics not geographically or seasonally structured.

Acknowledgments

We would like to thank members of the Konstantinidis Lab for sample collection and processing. This work was supported by US National Science Foundation [award 1241046]. DT was supported, in part, by an Onassis Foundation doctoral scholarship.

REFERENCES

Acinas, S. G., Klepac-Ceraj, V., Hunt, D. E., Pharino, C., Ceraj, I., Distel, D. L., & Polz, M. F. (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, *430*(6999), 551–554. https://doi.org/10.1038/nature02649

Allison, S. D., & Martiny, J. B. H. (2008). Resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences*, *105*(Supplement 1), 11512–11519. https://doi.org/10.1073/pnas.0801925105

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106. https://doi.org/10.1186/gb-2010-11-10-r106

Atlas, R. M., & Hazen, T. C. (2011). Oil Biodegradation and Bioremediation: A Tale of the Two Worst Spills in U.S. History. *Environmental Science & Technology*, *45*(16), 6709–6715. https://doi.org/10.1021/es2013227

Baas Becking, L. G. M. (1934). *Geobiologie of inleiding tot de milieukunde*. Den Haag: W.P. Van Stockum & Zoon.

Bahl, J., Lau, M. C. Y., Smith, G. J. D., Vijaykrishna, D., Cary, S. C., Lacap, D. C., … Pointing, S. B. (2011). Ancient origins determine global

biogeography of hot and cold desert cyanobacteria. *Nature Communications*, *2*, 163. https://doi.org/10.1038/ncomms1167

Balzer, S., Malde, K., Grohme, M. A., & Jonassen, I. (2013). Filtering duplicate reads from 454 pyrosequencing data. *Bioinformatics*, *29*(7), 830–836. https://doi.org/10.1093/bioinformatics/btt047

Barreto, D. P., Conrad, R., Klose, M., Claus, P., & Enrich-Prast, A. (2014). Distance-Decay and Taxa-Area Relationships for Bacteria, Archaea and Methanogenic Archaea in a Tropical Lake Sediment. *PLOS ONE*, *9*(10), e110128. https://doi.org/10.1371/journal.pone.0110128

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., … Lander, E. S. (2002). ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research*, *12*(1), 177–189. https://doi.org/10.1101/gr.208902

Bazzaz, F. A., & Pickett, S. T. A. (1980). Physiological Ecology of Tropical Succession: A Comparative Review. *Annual Review of Ecology and Systematics*, *11*(1), 287–310. https://doi.org/10.1146/annurev.es.11.110180.001443

Bengtsson, J., Eriksson, K. M., Hartmann, M., Wang, Z., Shenoy, B. D., Grelet, G.-A., … Nilsson, R. H. (2011). Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek*, *100*(3), 471–475. https://doi.org/10.1007/s10482-011-9598-6

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology*, *60*(3), 291–302. https://doi.org/10.1093/sysbio/syr010

Berthe-Corti, L., & Nachtkamp, M. (2010). Bacterial Communities in Hydrocarbon-Contaminated Marine Coastal Environments. In K. N. Timmis (Ed.), *Handbook of Hydrocarbon and Lipid Microbiology* (pp. 2349–2359). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/referenceworkentry/10.1007/978-3-540-77587-4_171

Bik, H. M., Halanych, K. M., Sharma, J., & Thomas, W. K. (2012). Dramatic Shifts in Benthic Microbial Eukaryote Communities following the Deepwater Horizon Oil Spill. *PLoS ONE*, *7*(6), e38550. https://doi.org/10.1371/journal.pone.0038550

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., … Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, *10*(1), 57–59. https://doi.org/10.1038/nmeth.2276

Boone, D. R., Castenholz, R. W., & Garrity, G. M. (Eds.). (2001). *Bergey's Manual® of Systematic Bacteriology*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-0-387-21609-6

Bouck, J., Miller, W., Gorrell, J. H., Muzny, D., & Gibbs, R. A. (1998). Analysis of the Quality and Utility of Random Shotgun Sequencing at Low Redundancies. *Genome Research*, *8*(10), 1074–1084. https://doi.org/10.1101/gr.8.10.1074

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, *6*(9), 673–676. https://doi.org/10.1038/nmeth.1358

Branscomb, E., & Predki, P. (2002). On the High Value of Low Standards. *Journal of Bacteriology*, *184*(23), 6406–6409. https://doi.org/10.1128/JB.184.23.6406-6409.2002

Brendan Logue, J., & Lindström, E. S. (2008). Biogeography of Bacterioplankton in Inland Waters. *Freshwater Reviews*, *1*(1), 99–114. https://doi.org/10.1608/FRJ-1.1.9

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*, 94. https://doi.org/10.1186/1471-2105-11-94

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., … Jaffe, D. B. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Research*, *18*(5), 810–820. https://doi.org/10.1101/gr.7337908

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Caporaso, J. G., Paszkiewicz, K., Field, D., Knight, R., & Gilbert, J. A. (2012). The Western English Channel contains a persistent microbial seed bank. *The ISME Journal*, *6*(6), 1089–1093. https://doi.org/10.1038/ismej.2011.162

Caro-Quintero, A., & Konstantinidis, K. T. (2012). Bacterial species may exist, metagenomics reveal. *Environmental Microbiology*, *14*(2), 347–355. https://doi.org/10.1111/j.1462-2920.2011.02668.x

Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research*, *19*(2), 336–346. https://doi.org/10.1101/gr.079053.108

Chaisson, M. J., & Pevzner, P. A. (2008). Short read fragment assembly of bacterial genomes. *Genome Research*, *18*(2), 324–330. https://doi.org/10.1101/gr.7088808

Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, *11*(4), 265–270.

Chao, A., & Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, *10*(4), 429–443. https://doi.org/10.1023/A:1026096204727

Charuvaka, A., & Rangwala, H. (2011). Evaluation of short read metagenomic assembly. *BMC Genomics*, *12*(2), 1–13. https://doi.org/10.1186/1471-2164-12-S2-S8

Cho, J.-C., & Tiedje, J. M. (2000). Biogeography and Degree of Endemicity of Fluorescent Pseudomonas Strains in Soil. *Applied and Environmental Microbiology*, *66*(12), 5448–5456. https://doi.org/10.1128/AEM.66.12.5448-5456.2000

Chrismas, N. A. M., Anesio, A. M., & Sánchez-Baracaldo, P. (2015). Multiple adaptations to polar and alpine environments within cyanobacteria: a phylogenomic and Bayesian approach. *Extreme Microbiology*, 1070. https://doi.org/10.3389/fmicb.2015.01070

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … Hoon, M. J. L. de. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cole, J. R., Konstantinidis, K., Farris, R. J., & Tiedje, J. M. (2010). Microbial diversity and phylogeny: extending from rRNAs to genomes. In W.-T. Liu & J. K. Jansson, *Environmental Molecular Biology*. Norwich, UK: Horizon Scientific Press.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., … Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(D1), D633–D642. https://doi.org/10.1093/nar/gkt1244

Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, *11*, 485. https://doi.org/10.1186/1471-2105-11-485

Curtis, T. P., Sloan, W. T., & Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences*, *99*(16), 10494–10499. https://doi.org/10.1073/pnas.142680199

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods*, *10*(4), 325–327. https://doi.org/10.1038/nmeth.2375

Delmont, T. O., Simonet, P., & Vogel, T. M. (2012). Describing microbial communities and performing global comparisons in the 'omic era. *The ISME Journal*, *6*(9), 1625–1628. https://doi.org/10.1038/ismej.2012.55

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., … Karl, D. M. (2006). Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science*, *311*(5760), 496–503. https://doi.org/10.1126/science.1120250

Denef, V. J., & Banfield, J. F. (2012). In Situ Evolutionary Rate Measurements Show Ecological Success of Recently Emerged Bacterial Hybrids. *Science*, *336*(6080), 462–466. https://doi.org/10.1126/science.1218389

Dennis, J. E., Jr., Gay, D. M., & Walsh, R. E. (1981). An Adaptive Nonlinear Least-Squares Algorithm. *ACM Trans. Math. Softw.*, *7*(3), 348–368. https://doi.org/10.1145/355958.355965

Dethlefsen, L., & Schmidt, T. M. (2007). Performance of the Translational Apparatus Varies with the Ecological Strategies of Bacteria. *Journal of Bacteriology*, *189*(8), 3237–3245. https://doi.org/10.1128/JB.01686-06

Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K., & Nattkemper, T. W. (2009). TACOA – Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, *10*, 56. https://doi.org/10.1186/1471-2105-10-56

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2007). SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Research*, *17*(11), 1697–1706. https://doi.org/10.1101/gr.6435207

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105–e105. https://doi.org/10.1093/nar/gkn425

Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M.-J., Richter, R. A., Valas, R., … Venter, J. C. (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *The ISME Journal*, *6*(6), 1186–1199. https://doi.org/10.1038/ismej.2011.189

Ecma International. (2013, October). The JSON Data Interchange Format. Ecma International. Retrieved from http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, *323*(5910), 133–138. https://doi.org/10.1126/science.1162986

Eppley, J. M., Tyson, G. W., Getz, W. M., & Banfield, J. F. (2007). Strainer: software for analysis of population variation in community genomic datasets. *BMC Bioinformatics*, *8*, 398. https://doi.org/10.1186/1471-2105-8-398

Esri Hydrology Team. (2012, May 16). World Hydro Reference Overlay. Retrieved October 19, 2016, from

http://www.arcgis.com/home/item.html?id=f7c73101a09c44058f8f029eefd37bd6

Esri Inc. (2015, March 16). World Land Cover 30m BaseVue 2013. Retrieved October 19, 2016, from http://www.arcgis.com/home/item.html?id=1770449f11df418db482a14df4ac26eb

Esty, W. W. (1986). The Efficiency of Good's Nonparametric Coverage Estimator. *The Annals of Statistics*, *14*(3), 1257–1260.

Fang, Z., Martin, J., & Wang, Z. (2012). Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell & Bioscience*, *2*, 26. https://doi.org/10.1186/2045-3701-2-26

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., … Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, *109*(52), 21390–21395. https://doi.org/10.1073/pnas.1215210110

Fierer, N., & Lennon, J. T. (2011). The generation and maintenance of diversity in microbial communities. *American Journal of Botany*, *98*(3), 439–448. https://doi.org/10.3732/ajb.1000498

Fierer, N., Morse, J. L., Berthrong, S. T., Bernhardt, E. S., & Jackson, R. B. (2007). Environmental Controls on the Landscape-Scale Biogeography of Stream Bacterial Communities. *Ecology*, *88*(9), 2162–2173. https://doi.org/10.1890/06-1746.1

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl 2), W29–W37. https://doi.org/10.1093/nar/gkr367

Fraser, C., Hanage, W. P., & Spratt, B. G. (2007). Recombination and the Nature of Bacterial Speciation. *Science*, *315*(5811), 476–480. https://doi.org/10.1126/science.1127573

Fuhrman, J. A., Cram, J. A., & Needham, D. M. (2015). Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, *13*(3), 133–146. https://doi.org/10.1038/nrmicro3417

Fuhrman, J. A., Hewson, I., Schwalbach, M. S., Steele, J. A., Brown, M. V., & Naeem, S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences*, *103*(35), 13104–13109. https://doi.org/10.1073/pnas.0602399103

Gauthier, M. J., Lafay, B., Christen, R., Fernandez, L., Acquaviva, M., Bonin, P., & Bertrand, J.-C. (1992). Marinobacter hydrocarbonoclasticus gen. nov., sp. nov., a New, Extremely Halotolerant, Hydrocarbon-Degrading Marine Bacterium. *International Journal of Systematic Bacteriology*, *42*(4), 568–576. https://doi.org/10.1099/00207713-42-4-568

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., … Swings, J. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, *3*(9), 733–739. https://doi.org/10.1038/nrmicro1236

Gibbons, S. M., Caporaso, J. G., Pirrung, M., Field, D., Knight, R., & Gilbert, J. A. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proceedings of the National Academy of Sciences*, *110*(12), 4651–4655. https://doi.org/10.1073/pnas.1217767110

Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., … Field, D. (2012). Defining seasonal marine microbial community dynamics. *The ISME Journal*, *6*(2), 298–308. https://doi.org/10.1038/ismej.2011.107

Giovannoni, S. J., & Vergin, K. L. (2012). Seasonality in Ocean Microbial Communities. *Science*, *335*(6069), 671–676. https://doi.org/10.1126/science.1198078

Glaeser, J., & Overmann, J. (2004). Biogeography, Evolution, and Diversity of Epibionts in Phototrophic Consortia. *Applied and Environmental*

Microbiology, *70*(8), 4821–4830. https://doi.org/10.1128/AEM.70.8.4821-4830.2004

Gogarten, J. P., & Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, *3*(9), 679–687. https://doi.org/10.1038/nrmicro1204

Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, *40*(3–4), 237–264. https://doi.org/10.1093/biomet/40.3-4.237

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*(1), 81–91. https://doi.org/10.1099/ijs.0.64483-0

Green, J., & Bohannan, B. J. M. (2006). Spatial scaling of microbial biodiversity. *Trends in Ecology & Evolution*, *21*(9), 501–507. https://doi.org/10.1016/j.tree.2006.06.012

Greer, C. W. (2010). Bacterial Diversity in Hydrocarbon-Polluted Rivers, Estuaries and Sediments. In K. N. Timmis (Ed.), *Handbook of Hydrocarbon and Lipid Microbiology* (pp. 2329–2338). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/referenceworkentry/10.1007/978-3-540-77587-4_169

Gucht, K. V. der, Cottenie, K., Muylaert, K., Vloemans, N., Cousin, S., Declerck, S., … Meester, L. D. (2007). The power of species sorting: Local factors drive bacterial community composition over a wide range of spatial scales. *Proceedings of the National Academy of Sciences*, *104*(51), 20404–20409. https://doi.org/10.1073/pnas.0707200104

Hallam, S. J., Konstantinidis, K. T., Putnam, N., Schleper, C., Watanabe, Y., Sugahara, J., … DeLong, E. F. (2006). Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum.

Proceedings of the National Academy of Sciences, *103*(48), 18296–18301. https://doi.org/10.1073/pnas.0608549103

Handelsman, J., Tiedje, J. M., Alvarez-Cohen, L., Ashburner, M., Cann, I. K. O., DeLong, E. F., … Schmid, M. B. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/11902

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., & Martiny, J. B. H. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*, *10*(7), 497–506. https://doi.org/10.1038/nrmicro2795

Hassan, A., Taqi, Z., Obuekwe, C., & Al-Saleh, E. (2011). Characterization of crude oil-degrading bacteria in a crude oil-contaminated and uncontaminated site in Kuwait (pp. 177–186). https://doi.org/10.2495/ST110161

Hausser, J., & Strimer, K. (2013). entropy: Estimation of Entropy, Mutual Information and Related Quantities (Version 1.2.0). Retrieved from http://cran.r-project.org/web/packages/entropy/index.html

Hayworth, J. S., Clement, T. P., & Valentine, J. F. (2011). Deepwater Horizon oil spill impacts on Alabama beaches. *Hydrology and Earth System Sciences Discussions*, *8*(4), 6721–6747. https://doi.org/10.5194/hessd-8-6721-2011

Hazen, T. C., Dubinsky, E. A., DeSantis, T. Z., Andersen, G. L., Piceno, Y. M., Singh, N., … Mason, O. U. (2010). Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria. *Science*, *330*(6001), 204–208. https://doi.org/10.1126/science.1195979

Head, I. M., Jones, D. M., & Röling, W. F. M. (2006). Marine microorganisms make a meal of oil. *Nature Reviews Microbiology*, *4*(3), 173–182. https://doi.org/10.1038/nrmicro1348

Hooper, S. D., Dalevi, D., Pati, A., Mavromatis, K., Ivanova, N. N., & Kyrpides, N. C. (2010). Estimating DNA coverage and abundance in metagenomes

using a gamma approximation. *Bioinformatics*, *26*(3), 295–301. https://doi.org/10.1093/bioinformatics/btp687

Horton, M., Bodenhausen, N., & Bergelson, J. (2010). MARTA: a suite of Java-based tools for assigning taxonomic status to DNA sequences. *Bioinformatics*, *26*(4), 568–569. https://doi.org/10.1093/bioinformatics/btp682

Hubbell, S. P. (2008). *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32).* Princeton: Princeton University Press. Retrieved from http://public.eblib.com/choice/publicfullrecord.aspx?p=714075

Huettel, M., Berg, P., & Kostka, J. E. (2014). Benthic Exchange and Biogeochemical Cycling in Permeable Sediments. *Annual Review of Marine Science*, *6*(1), 23–51. https://doi.org/10.1146/annurev-marine-051413-012706

Hugenholtz, P., & Tyson, G. W. (2008). Microbiology: Metagenomics. *Nature*, *455*(7212), 481–483. https://doi.org/10.1038/455481a

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., & Bohannan, B. J. M. (2001). Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. *Applied and Environmental Microbiology*, *67*(10), 4399–4406. https://doi.org/10.1128/AEM.67.10.4399-4406.2001

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Iverson, V., Morris, R. M., Frazar, C. D., Berthiaume, C. T., Morales, R. L., & Armbrust, E. V. (2012). Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science*, *335*(6068), 587–590. https://doi.org/10.1126/science.1212665

Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., … Jones, C. D. (2007). Extending assembly of short DNA sequences to handle error. *Bioinformatics*, *23*(21), 2942–2944. https://doi.org/10.1093/bioinformatics/btm451

Jones, S. E., Cadkin, T. A., Newton, R. J., & McMahon, K. D. (2012). Spatial and temporal scales of aquatic bacterial beta diversity. *Aquatic Microbiology*, *3*, 318. https://doi.org/10.3389/fmicb.2012.00318

Jones, S. E., & McMahon, K. D. (2009). Species-sorting may explain an apparent minimal effect of immigration on freshwater bacterial community dynamics. *Environmental Microbiology*, *11*(4), 905–913. https://doi.org/10.1111/j.1462-2920.2008.01814.x

Joye, S. B., Teske, A. P., & Kostka, J. E. (2014). Microbial Dynamics Following the Macondo Oil Well Blowout across Gulf of Mexico Environments. *BioScience*, *64*(9), 766–777. https://doi.org/10.1093/biosci/biu121

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Kent, A. D., Yannarell, A. C., Rusak, J. A., Triplett, E. W., & McMahon, K. D. (2007). Synchrony in aquatic microbial community dynamics. *The ISME Journal*, *1*(1), 38–47. https://doi.org/10.1038/ismej.2007.6

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kertesz, M. A., & Kawasaki, A. (2010). Hydrocarbon-Degrading Sphingomonads: Sphingomonas, Sphingobium, Novosphingobium, and Sphingopyxis. In K. N. Timmis (Ed.), *Handbook of Hydrocarbon and Lipid Microbiology* (pp. 1693–1705). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/referenceworkentry/10.1007/978-3-540-77587-4_119

King, G. M., Kostka, J. E., Hazen, T. C., & Sobecky, P. A. (2015). Microbial Responses to the Deepwater Horizon Oil Spill in the Northern Gulf of

Mexico: From Coastal Wetlands to the Deep Sea. *Annual Review of Marine Science*, *7*. https://doi.org/10.1146/annurev-marine-010814-015543

Konstantinidis, K. T., Braff, J., Karl, D. M., & DeLong, E. F. (2009). Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North Pacific Subtropical Gyre. *Applied and Environmental Microbiology*, *75*(16), 5345–5355. https://doi.org/10.1128/AEM.00473-09

Konstantinidis, K. T., & DeLong, E. F. (2008). Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *The ISME Journal*, *2*(10). https://doi.org/10.1038/ismej.2008.62

Konstantinidis, K. T., Ramette, A., & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1475), 1929–1940. https://doi.org/10.1098/rstb.2006.1920

Konstantinidis, K. T., & Stackebrandt, E. (2013). Defıning taxonomic ranks. In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, & E. Stackebrandt (Eds.), *The Prokaryotes* (Vol. 1, pp. 29–57). New York, NY: Springer-Verlag.

Konstantinidis, K. T., & Tiedje, J. M. (2005a). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2567–2572. https://doi.org/10.1073/pnas.0409727102

Konstantinidis, K. T., & Tiedje, J. M. (2005b). Towards a Genome-Based Taxonomy for Prokaryotes. *Journal of Bacteriology*, *187*(18), 6258–6264. https://doi.org/10.1128/JB.187.18.6258-6264.2005

Konstantinidis, K. T., & Tiedje, J. M. (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology*, *10*(5), 504–509. https://doi.org/10.1016/j.mib.2007.08.006

Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., … Gerstein, M. B. (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, *10*, R23. https://doi.org/10.1186/gb-2009-10-2-r23

Kostka, J. E., Prakash, O., Overholt, W. A., Green, S. J., Freyer, G., Canion, A., … Huettel, M. (2011). Hydrocarbon-Degrading Bacteria and the Bacterial Community Response in Gulf of Mexico Beach Sands Impacted by the Deepwater Horizon Oil Spill. *Applied and Environmental Microbiology*, *77*(22), 7962–7974. https://doi.org/10.1128/AEM.05402-11

Kostka, J. E., Teske, A., Joye, S. B., & Head, I. M. (2014). The metabolic pathways and environmental controls of hydrocarbon biodegradation in marine ecosystems. *Frontiers in Microbiology*, *5*, 471. https://doi.org/10.3389/fmicb.2014.00471

Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., … Stoye, J. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research*, *36*(7), 2230–2239. https://doi.org/10.1093/nar/gkn038

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews*, *72*(4), 557–578. https://doi.org/10.1128/MMBR.00009-08

Kunin, V., Engelbrektson, A., Ochman, H., & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, *12*(1), 118–123. https://doi.org/10.1111/j.1462-2920.2009.02051.x

Lai, Q., Wang, L., Liu, Y., Yuan, J., Sun, F., & Shao, Z. (2011). Parvibaculum indicum sp. nov., isolated from deep-sea water. *International Journal of Systematic and Evolutionary Microbiology*, *61*(Pt 2), 271–274. https://doi.org/10.1099/ijs.0.021899-0

Lamendella, R., Strutt, S., Borglin, S. E., Chakraborty, R., Tas, N., Mason, O. U., … Jansson, J. (2014). Assessment of the Deepwater Horizon oil spill impact on Gulf coast microbial communities. *Aquatic Microbiology*, *5*, 130. https://doi.org/10.3389/fmicb.2014.00130

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, *2*(3), 231–239. https://doi.org/10.1016/0888-7543(88)90007-9

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lehner, B., Verdin, K., & Jarvis, A. (2008). New Global Hydrography Derived From Spaceborne Elevation Data. *Eos, Transactions American Geophysical Union*, *89*(10), 93–94. https://doi.org/10.1029/2008EO100001

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128. https://doi.org/10.1093/bioinformatics/btl529

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. https://doi.org/10.1101/gr.078212.108

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., … Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265–272. https://doi.org/10.1101/gr.097261.109

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., … Lopez, R. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Research*, *43*(W1), W580–W584. https://doi.org/10.1093/nar/gkv279

Lindström, E. S., & Langenheder, S. (2012). Local and regional factors influencing bacterial community assembly: Bacterial community assembly. *Environmental Microbiology Reports*, *4*(1), 1–9. https://doi.org/10.1111/j.1758-2229.2011.00257.x

Liu, J., Wang, H., Yang, H., Zhang, Y., Wang, J., Zhao, F., & Qi, J. (2013). Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic Acids Research*, *41*(1), e3–e3. https://doi.org/10.1093/nar/gks828

Lu, Z., Deng, Y., Van Nostrand, J. D., He, Z., Voordeckers, J., Zhou, A., … Zhou, J. (2012). Microbial gene functions enriched in the Deepwater Horizon deep-sea oil plume. *The ISME Journal*, *6*(2), 451–460. https://doi.org/10.1038/ismej.2011.91

Lubchenco, J., McNutt, M., Lehr, B., Sogge, M., Miller, M., Hammond, S., & Conner, W. (2010). *Deepwater Horizon/BP Oil Budget: What happened to the oil?* Retrieved from http://www.noaanews.noaa.gov/stories2010/PDFs/OilBudget_description_%2083final.pdf

Luis M Rodriguez-R, & Konstantinos T Konstantinidis. (2016). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. https://doi.org/10.7287/peerj.preprints.1900v1

Luo, C., & Konstantinidis, K. T. (2011). Phosphorus-related gene content is similar in Prochlorococcus populations from the North Pacific and North Atlantic Oceans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(16), E62–E63. https://doi.org/10.1073/pnas.1018662108

Luo, C., Rodriguez-R, L. M., & Konstantinidis, K. T. (2014). MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research*, *42*(8), e73. https://doi.org/10.1093/nar/gku169

Luo, C., Tsementzi, D., Kyrpides, N. C., & Konstantinidis, K. T. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *The ISME Journal*, *6*(4), 898–901. https://doi.org/10.1038/ismej.2011.147

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., & Konstantinidis, K. T. (2012). Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLOS ONE*, *7*(2), e30087. https://doi.org/10.1371/journal.pone.0030087

Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., & Konstantinidis, K. T. (2011). Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences*, *108*(17), 7200–7205. https://doi.org/10.1073/pnas.1015622108

Mackelprang, R., Waldrop, M. P., DeAngelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., … Jansson, J. K. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, *480*(7377), 368–371. https://doi.org/10.1038/nature10576

Maeda, R., Ito, Y., Iwata, K., & Omori, T. (2010). Comparison of marine and terrestrial carbazole-degrading bacteria. *Curr Res Technol Educ Top Appl Microbiol Microb Biotechnol*, *2*, 1311–1321.

Maeda, R., Nagashima, H., Widada, J., Iwata, K., & Omori, T. (2009). Novel marine carbazole-degrading bacteria. *FEMS Microbiology Letters*, *292*(2), 203–209. https://doi.org/10.1111/j.1574-6968.2009.01497.x

Mande, S. S., Mohammed, M. H., & Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics*, *13*(6), 669–681. https://doi.org/10.1093/bib/bbs054

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., … Staley, J. T. (2006). Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, *4*(2), 102–112. https://doi.org/10.1038/nrmicro1341

Matsen, F. A., Hoffman, N. G., Gallagher, A., & Stamatakis, A. (2012). A Format for Phylogenetic Placements. *PLoS ONE*, *7*(2), e31009. https://doi.org/10.1371/journal.pone.0031009

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. https://doi.org/10.1186/1471-2105-11-538

McGenity, T. J., Folwell, B. D., McKew, B. A., & Sanni, G. O. (2012). Marine crude-oil biodegradation: a central role for interspecies interactions. *Aquatic Biosystems*, *8*(1), 10. https://doi.org/10.1186/2046-9063-8-10

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, *4*(1), 63–72. https://doi.org/10.1038/nmeth976

McMurdie, P. J., & Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Comput Biol*, *10*(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., … Edwards, R. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, *9*, 386. https://doi.org/10.1186/1471-2105-9-386

Michel, J., Owens, E. H., Zengel, S., Graham, A., Nixon, Z., Allard, T., … Taylor, E. (2013). Extent and Degree of Shoreline Oiling: Deepwater Horizon Oil Spill, Gulf of Mexico, USA. *PLoS ONE*, *8*(6), e65087. https://doi.org/10.1371/journal.pone.0065087

Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W., & Banfield, J. F. (2011). EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biology*, *12*, R44. https://doi.org/10.1186/gb-2011-12-5-r44

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Morowitz, M. J., Denef, V. J., Costello, E. K., Thomas, B. C., Poroyko, V., Relman, D. A., & Banfield, J. F. (2011). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proceedings of the National Academy of Sciences*, *108*(3), 1128–1133. https://doi.org/10.1073/pnas.1010992108

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., … Venter, J. C. (2000). A Whole-Genome Assembly of Drosophila. *Science*, *287*(5461), 2196–2204. https://doi.org/10.1126/science.287.5461.2196

Nalbantoglu, O. U., Way, S. F., Hinrichs, S. H., & Sayood, K. (2011). RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics*, *12*, 41. https://doi.org/10.1186/1471-2105-12-41

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155–e155. https://doi.org/10.1093/nar/gks678

Nemergut, D. R., Costello, E. K., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S. K., … Knight, R. (2011). Global patterns in the biogeography of bacterial taxa: Global bacterial biogeography. *Environmental Microbiology*, *13*(1), 135–144. https://doi.org/10.1111/j.1462-2920.2010.02315.x

Newton, R. J., Huse, S. M., Morrison, H. G., Peake, C. S., Sogin, M. L., & McLellan, S. L. (2013). Shifts in the Microbial Community Composition of

Gulf Coast Beaches Following Beach Oiling. *PLoS ONE*, *8*(9), e74265. https://doi.org/10.1371/journal.pone.0074265

Newton, R. J., Jones, S. E., Helmus, M. R., & McMahon, K. D. (2007). Phylogenetic Ecology of the Freshwater Actinobacteria acI Lineage. *Applied and Environmental Microbiology*, *73*(22), 7169–7176. https://doi.org/10.1128/AEM.00794-07

Oda, Y., Star, B., Huisman, L. A., Gottschal, J. C., & Forney, L. J. (2003). Biogeography of the Purple Nonsulfur Bacterium Rhodopseudomonas palustris. *Applied and Environmental Microbiology*, *69*(9), 5186–5191. https://doi.org/10.1128/AEM.69.9.5186-5191.2003

Ofiţeru, I. D., Lunn, M., Curtis, T. P., Wells, G. F., Criddle, C. S., Francis, C. A., & Sloan, W. T. (2010). Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences*, *107*(35), 15345–15350. https://doi.org/10.1073/pnas.1000604107

Oh, S., Caro-Quintero, A., Tsementzi, D., DeLeon-Rodriguez, N., Luo, C., Poretsky, R., & Konstantinidis, K. T. (2011). Metagenomic Insights into the Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Applied and Environmental Microbiology*, *77*(17), 6000–6011. https://doi.org/10.1128/AEM.00107-11

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*, 132. https://doi.org/10.1186/s13059-016-0997-x

Parks, D. H., & Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, *26*(6), 715–721. https://doi.org/10.1093/bioinformatics/btq041

Patil, K. R., Haider, P., Pope, P. B., Turnbaugh, P. J., Morrison, M., Scheffer, T., & McHardy, A. C. (2011). Taxonomic metagenome sequence assignment

with structured output models. *Nature Methods*, *8*(3), 191–192. https://doi.org/10.1038/nmeth0311-191

Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, *10*(12), 1200–1202. https://doi.org/10.1038/nmeth.2658

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., & Brown, C. T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences*, *109*(33), 13272–13277. https://doi.org/10.1073/pnas.1121464109

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*, *27*(13), i94–i101. https://doi.org/10.1093/bioinformatics/btr216

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Pintor, L. M., Brown, J. S., & Vincent, T. L. (2011). Evolutionary Game Theory as a Framework for Studying Biological Invasions. *The American Naturalist*, *177*(4), 410–423. https://doi.org/10.1086/658149

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS ONE*, *9*(4), e93827. https://doi.org/10.1371/journal.pone.0093827

Preston, F. W. (1948). The Commonness, And Rarity, of Species. *Ecology*, *29*(3), 254–283. https://doi.org/10.2307/1930989

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., … Young, J. P. W. (2007). The role of ecological theory in microbial ecology. *Nature Reviews. Microbiology*, *5*(5), 384–392. https://doi.org/10.1038/nrmicro1643

Ramette, A., & Tiedje, J. M. (2006). Biogeography: An Emerging Cornerstone for Understanding Prokaryotic Diversity, Ecology, and Evolution. *Microbial Ecology*, *53*(2), 197–207. https://doi.org/10.1007/s00248-005-5010-2

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., … Forney, L. J. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, *108*(Supplement 1), 4680–4687. https://doi.org/10.1073/pnas.1002611107

Reed, H. E., & Martiny, J. B. H. (2007). Testing the functional significance of microbial composition in natural communities. *FEMS Microbiology Ecology*, *62*(2), 161–170. https://doi.org/10.1111/j.1574-6941.2007.00386.x

Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, *38*(20), e191. https://doi.org/10.1093/nar/gkq747

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Riesenfeld, C. S., Schloss, P. D., & Handelsman, J. (2004). Metagenomics: Genomic Analysis of Microbial Communities. *Annual Review of Genetics*, *38*(1), 525–552. https://doi.org/10.1146/annurev.genet.38.072902.091216

Rodriguez-R, L. M., Castro, J. C., & Konstantinidis, K. T. (In preparation). How much rRNA gene surveys underestimate extant microbial diversity?

Rodriguez-R, L. M., Grajales, A., Arrieta-Ortiz, M., Salazar, C., Restrepo, S., & Bernal, A. (2012). Genomes-based phylogeny of the genus Xanthomonas. *BMC Microbiology*, *12*(1), 43. https://doi.org/10.1186/1471-2180-12-43

Rodriguez-R, L. M., Gunturu, S., Guo, J., Luo, C., Tiedje, J. M., Cole, J. R., & Konstantinidis, K. T. (Under review). Nonpareil 3: rapid estimation of metagenomic coverage and sequence diversity.

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014a). Bypassing Cultivation To Identify Bacterial Species. *Microbe*, *9*(3), 111–118.

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014b). Estimating coverage in metagenomic data sets and why it matters. *The ISME Journal*, *8*(11), 2349–2351. https://doi.org/10.1038/ismej.2014.76

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014c). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics*, *30*(5), 629–635. https://doi.org/10.1093/bioinformatics/btt584

Rodriguez-R, L. M., Overholt, W. A., Hagan, C., Huettel, M., Kostka, J. E., & Konstantinidis, K. T. (2015). Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. *The ISME Journal*. https://doi.org/10.1038/ismej.2015.5

Röling, W. F. M., Milner, M. G., Jones, D. M., Fratepietro, F., Swannell, R. P. J., Daniel, F., & Head, I. M. (2004). Bacterial Community Dynamics and Hydrocarbon Degradation during a Field-Scale Evaluation of Bioremediation on a Mudflat Beach Contaminated with Buried Oil. *Applied and Environmental Microbiology*, *70*(5), 2603–2613. https://doi.org/10.1128/AEM.70.5.2603-2613.2004

Röling, W. F. M., Milner, M. G., Jones, D. M., Lee, K., Daniel, F., Swannell, R. J. P., & Head, I. M. (2002). Robust Hydrocarbon Degradation and Dynamics of Bacterial Communities during Nutrient-Enhanced Oil Spill Bioremediation. *Applied and Environmental Microbiology*, *68*(11), 5537–5548. https://doi.org/10.1128/AEM.68.11.5537-5548.2002

Röling, W. F. M., & van Bodegom, P. M. (2014). Toward quantitative understanding on microbial community structure and functioning: a modeling-centered approach using degradation of marine oil spills as example. *Aquatic Microbiology*, *5*, 125. https://doi.org/10.3389/fmicb.2014.00125

Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011). NBC: the Naïve Bayes Classification tool webserver for taxonomic classification of

metagenomic reads. *Bioinformatics*, *27*(1), 127–129. https://doi.org/10.1093/bioinformatics/btq619

Rosselló-Mora, R., & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiology Reviews*, *25*(1), 39–67. https://doi.org/10.1016/S0168-6445(00)00040-1

Rousk, J., Bååth, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., … Fierer, N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *The ISME Journal*, *4*(10), 1340–1351. https://doi.org/10.1038/ismej.2010.58

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., … Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *37*(Database issue), D5-15. https://doi.org/10.1093/nar/gkn741

Schloss, P. D., & Handelsman, J. (2005). Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology*, *71*(3), 1501–1506. https://doi.org/10.1128/AEM.71.3.1501-1506.2005

Schloss, P. D., & Handelsman, J. (2008). A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics*, *9*, 34. https://doi.org/10.1186/1471-2105-9-34

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. https://doi.org/10.1038/nmeth.2089

Schneiker, S., dos Santos, V. A. M., Bartels, D., Bekel, T., Brecht, M., Buhrmester, J., … Golyshin, P. N. (2006). Genome sequence of the

ubiquitous hydrocarbon-degrading marine bacterium Alcanivorax borkumensis. *Nature Biotechnology*, *24*(8), 997–1004. https://doi.org/10.1038/nbt1232

Schreiber, F., Gumrich, P., Daniel, R., & Meinicke, P. (2010). Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, *26*(7), 960–961. https://doi.org/10.1093/bioinformatics/btq070

Shade, A., Peter, H., Allison, S. D., Baho, D. L., Berga, M., Bürgmann, H., … Handelsman, J. (2012). Fundamentals of microbial community resistance and resilience. *Frontiers in Microbiology*, *3*, 417. https://doi.org/10.3389/fmicb.2012.00417

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., … Alm, E. J. (2012). Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*, *336*(6077), 48–51. https://doi.org/10.1126/science.1218198

Shea, K., & Chesson, P. (2002). Community ecology theory as a framework for biological invasions. *Trends in Ecology & Evolution*, *17*(4), 170–176. https://doi.org/10.1016/S0169-5347(02)02495-3

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1). https://doi.org/10.1038/msb.2011.75

Simmons, S. L., DiBartolo, G., Denef, V. J., Goltsman, D. S. A., Thelen, M. P., & Banfield, J. F. (2008). Population Genomic Analysis of Strain Variation in Leptospirillum Group II Bacteria Involved in Acid Mine Drainage Formation. *PLOS Biol*, *6*(7), e177. https://doi.org/10.1371/journal.pbio.0060177

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. https://doi.org/10.1101/gr.089532.108

Smith, C. B., Tolar, B. B., Hollibaugh, J. T., & King, G. M. (2013). Alkane hydroxylase gene (alkB) phylotype composition and diversity in northern Gulf of Mexico bacterioplankton. *Aquatic Microbiology*, *4*, 370. https://doi.org/10.3389/fmicb.2013.00370

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., … Birney, E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, *12*(10), 1611–1618. https://doi.org/10.1101/gr.361602

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, *22*(21), 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stanhope, S. A. (2010). Occupancy Modeling, Maximum Contig Size Probabilities and Designing Metagenomics Experiments. *PLOS ONE*, *5*(7), e11652. https://doi.org/10.1371/journal.pone.0011652

Stegen, J. C., Lin, X., Fredrickson, J. K., Chen, X., Kennedy, D. W., Murray, C. J., … Konopka, A. (2013). Quantifying community assembly processes and identifying features that impose them. *The ISME Journal*, *7*(11), 2069–2079. https://doi.org/10.1038/ismej.2013.93

Stepanauskas, R. (2012). Single cell genomics: an individual look at microbes. *Current Opinion in Microbiology*, *15*(5), 613–620. https://doi.org/10.1016/j.mib.2012.09.001

Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., & Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences*, *106*(19), 7702–7707. https://doi.org/10.1073/pnas.0901054106

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, *23*(10), 1282–1288. https://doi.org/10.1093/bioinformatics/btm098

Tamames, J., de la Peña, S., & de Lorenzo, V. (2012). COVER: a priori estimation of coverage for metagenomic sequencing. *Environmental Microbiology Reports*, *4*(3), 335–341. https://doi.org/10.1111/j.1758-2229.2012.00338.x

The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., … Pop, M. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biology*, *14*, R2. https://doi.org/10.1186/gb-2013-14-1-r2

Tsementzi, D., Poretsky, R., Rodriguez-R, L. M., Luo, C., & Konstantinidis, K. T. (2014). Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environmental Microbiology Reports*, *6*(6), 640–655. https://doi.org/10.1111/1758-2229.12180

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, *98*(9), 5116–5121. https://doi.org/10.1073/pnas.091062498

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., … Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43. https://doi.org/10.1038/nature02340

Valentine, D. L., Mezić, I., Maćešić, S., Črnjarić-Žic, N., Ivić, S., Hogan, P. J., … Loire, S. (2012). Dynamic autoinoculation and the microbial ecology of a deep water hydrocarbon irruption. *Proceedings of the National Academy*

of Sciences, *109*(50), 20286–20291. https://doi.org/10.1073/pnas.1108820109

Vázquez, D. P., & Simberloff, D. (2002). Ecological Specialization and Susceptibility to Disturbance: Conjectures and Refutations. *The American Naturalist*, *159*(6), 606–623. https://doi.org/10.1086/339991

Větrovský, T., & Baldrian, P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLOS ONE*, *8*(2), e57923. https://doi.org/10.1371/journal.pone.0057923

Vieira-Silva, S., & Rocha, E. P. C. (2010). The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLoS Genet*, *6*(1), e1000808. https://doi.org/10.1371/journal.pgen.1000808

Wang, L., Wang, W., Lai, Q., & Shao, Z. (2010). Gene diversity of CYP153A and AlkB alkane hydroxylases in oil-degrading bacteria isolated from the Atlantic Ocean. *Environmental Microbiology*, *12*(5), 1230–1242. https://doi.org/10.1111/j.1462-2920.2010.02165.x

Wang, P., & Roberts, T. M. (2013). Distribution of Surficial and Buried Oil Contaminants across Sandy Beaches along NW Florida and Alabama Coasts Following the Deepwater Horizon Oil Spill in 2010. *Journal of Coastal Research*, *291*, 144–155. https://doi.org/10.2112/JCOASTRES-D-12-00198.1

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Warren, R. L., Sutton, G. G., Jones, S. J. M., & Holt, R. A. (2007). Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, *23*(4), 500–501. https://doi.org/10.1093/bioinformatics/btl629

Weiher, E., Freund, D., Bunton, T., Stefanski, A., Lee, T., & Bentivenga, S. (2011). Advances, challenges and a developing synthesis of ecological

community assembly theory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *366*(1576), 2403–2413. https://doi.org/10.1098/rstb.2011.0056

Wendl, M. C. (2006). A General Coverage Theory for Shotgun DNA Sequencing. *Journal of Computational Biology*, *13*(6), 1177–1196. https://doi.org/10.1089/cmb.2006.13.1177

Wendl, M. C., Kota, K., Weinstock, G. M., & Mitreva, M. (2012). Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. *Journal of Mathematical Biology*, *67*(5), 1141–1161. https://doi.org/10.1007/s00285-012-0586-x

Wendl, M. C., Marra, M. A., Hillier, L. W., Chinwalla, A. T., Wilson, R. K., & Waterston, R. H. (2001). Theories and Applications for Sequencing Randomly Selected Clones. *Genome Research*, *11*(2), 274–280. https://doi.org/10.1101/gr.133901

White, J. R., Nagarajan, N., & Pop, M. (2009). Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLOS Comput Biol*, *5*(4), e1000352. https://doi.org/10.1371/journal.pcbi.1000352

Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E. M., Kyrpides, N., … Meyer, F. (2012). The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics*, *13*(1), 141. https://doi.org/10.1186/1471-2105-13-141

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Wolf, Y. I., & Koonin, E. V. (2012). A Tight Link between Orthologs and Bidirectional Best Hits in Bacterial and Archaeal Genomes. *Genome Biology and Evolution*, *4*(12), 1286–1294. https://doi.org/10.1093/gbe/evs100

Wren, J. D. (2004). 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*, *20*(5), 668–672. https://doi.org/10.1093/bioinformatics/btg465

Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., … Banfield, J. F. (2012). Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science*, *337*(6102), 1661–1665. https://doi.org/10.1126/science.1224041

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., … Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, *462*(7276), 1056–1060. https://doi.org/10.1038/nature08656

Wu, M., & Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, *9*, R151. https://doi.org/10.1186/gb-2008-9-10-r151

Xiong, J., Liu, Y., Lin, X., Zhang, H., Zeng, J., Hou, J., … Chu, H. (2012). Geographic distance and pH drive bacterial distribution in alkaline lake sediments across Tibetan Plateau. *Environmental Microbiology*, *14*(9), 2457–2466. https://doi.org/10.1111/j.1462-2920.2012.02799.x

Yakimov, M. M., Timmis, K. N., & Golyshin, P. N. (2007). Obligate oil-degrading marine bacteria. *Current Opinion in Biotechnology*, *18*(3), 257–266. https://doi.org/10.1016/j.copbio.2007.04.006

Yannarell, A. C., & Triplett, E. W. (2005). Geographic and Environmental Sources of Variation in Lake Bacterial Community Composition. *Applied and Environmental Microbiology*, *71*(1), 227–239. https://doi.org/10.1128/AEM.71.1.227-239.2005

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zhu, W., Lomsadze, A., & Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research*, *38*(12), e132. https://doi.org/10.1093/nar/gkq275

Zinder, S. H., & Salyers, A. A. (2015). Microbial Ecology—New Directions, New Importance. In *Bergey's Manual of Systematics of Archaea and Bacteria*. John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118960608.bm00015/abstract

Zuijdgeest, A., & Huettel, M. (2012). Dispersants as Used in Response to the MC252-Spill Lead to Higher Mobility of Polycyclic Aromatic Hydrocarbons in Oil-Contaminated Gulf of Mexico Sand. *PLoS ONE*, *7*(11), e50549. https://doi.org/10.1371/journal.pone.0050549

APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 6

A.1. Supplementary Methods

A.1.1. Redundancy Estimation Using K-mers

Nonpareil with k-mer kernel estimates the abundance-weighted average coverage of a metagenome sequence (combined genomes of all organisms in a sample) by a metagenomic dataset (a subset of the metagenome sequence obtained by shotgun sequencing of DNA isolated from the sample) by measuring the coverage of a subset of the positions in the metagenome sequence (target positions). To determine if a read from the dataset covers a target position, we require that the read match the k-mer sequence starting with the target position and including the next $k - 1$ bases in the metagenome sequence, with k chosen to be large enough that random matches in the metagenomic dataset will be uncommon. Note that this test is unable to detect if any of the last $k - 1$ bases in a metagenomic sequencing read cover a target position, so for each metagenomic read of length L, only $L - k + 1$ positions can be tested for matches, reducing the effective size of the metagenomic dataset scanned with respect to the complete alignment kernel.

The target k-mers are derived from a randomly selected subset of the metagenomic dataset reads, with one target k-mer selected from each read. Counting the number of times each target k-mer (or its reverse complement) is covered can be performed in time proportional to the size of the metagenomic dataset and is independent of the number of target k-mers. Because the target k-

mers are derived from the metagenomic sequence reads, some of the k-mers will contain sequence errors, but if *k* is large enough, these error k-mers will likely have zero coverage. Although it is not possible to determine which of the zero-coverage target k-mers contain errors, we utilize the sequencing quality scores (Q-scores) from each target k-mer's parent sequencing read to estimate *E*, the number of k-mers with errors. To correct for these errors, we reduce the target set by removing *E* zero-coverage target k-mers. The coverage counts for the remaining target k-mers, along with the reduced effective metagenomic dataset size, are passed through the remaining original Nonpareil steps to estimate required sequencing effort.

The original Nonpareil method (alignment kernel) measures read redundancy, as opposed to coverage, by comparing each metagenomic dataset read for sequence overlap with a small subset of target reads. When the minimum required overlap is set to 50%, this is essentially equivalent to directly measuring the coverage of the position at the center of each target read (target position). This target position is covered by any overlap of 50% or longer starting at either end of the target read and will not be covered by any shorter overlaps. Nonpareil 3 provides different methods to tune the amount of sequence similarity required to consider a target position "covered" on each kernel (alignment or k-mer). The alignment kernel includes a sequence similarity threshold for the overlap region while the k-mer kernel allows the k-mer length to be specified. From our tests, the k-mer kernel with a k-mer value of 24 provides similar estimated required sequencing efforts to the alignment kernel with a 95% identity threshold (§A.1.2 below).

The runtime of Nonpareil with alignment kernel is approximately proportional to $N \times L^2 \times T$, where N is the number of sequence reads, T is the number of target positions, and L the length of a read, while the complexity of Nonpareil with k-mer kernel is simply $N \times L$, the total dataset size. As the runtime for the k-mer kernel is not otherwise sensitive to read length, Nonpareil 3 is

directly compatible with long-read data (*e.g.*, PacBio or MinION). In addition, since the runtime is not proportional to *T*, it is practical to use a larger value for *T* (default 10,000, instead of 1,000 in the alignment kernel), increasing Nonpareil precision over the original version.

Nonpareil is implemented in C++ and was tested on both MacOS 10.11 and Linux Red Hat 4.4.7. The Nonpareil R script used for model fitting was modified to accept output of both Nonpareil k-mer and alignment kernels.

A.1.2. Testing Kernel Consistency

Metagenomic datasets were processed using Scythe (https://github.com/vsbuffalo/scythe) for adapter trimming. Then datasets were quality-trimmed using Solexa QA (Cox et al., 2010) with maximum expected error of 1% and minimum length of 50 bp. For paired-end samples, only the forward reads were used. Short read archive (SRA) identifiers are given for all the datasets except for Iowa Continuous Corn soil (Suppl. Table 1). For the latter, seven lanes from one run of Illumina HiSeq from raw data from project 402461 were retrieved as seven files named 1424.1.1371.fastq.gz to 1424.7.1371.fastq.gz from the JGI Genome Portal (http://genome.jgi.doe.gov follow download link; raw data folder) on July 21, 2016.

Nonpareil results with k-mer kernel were collected using a 27-inch iMac with an Intel core i5 3.2 GHz processor using k-mer size of 24, 10,000 queries, and default value of 2 threads (the k-mer matching portion of Nonpareil is single threaded). Nonpareil alignment kernel results for Iowa soil were collected using the Michigan State University High-Performance Computing Center (MSU HPCC) nodes with 20 cores using options: 20 threads, 1,000 queries, 50% overlap and using a redundancy to coverage transformation factor of 1.0. For comparing time between the methods, all other datasets were processed using a 27-inch iMac as above and with the same Nonpareil alignment options at 95% identity, except

using the default of 2 threads. Iowa soil CPU time for alignment kernel was estimated using the linear regression of the other datasets.

A.2. Supplementary Results

A.2.1. K-mer kernel testing

The Nonpareil k-mer kernel was compared to the alignment kernel (95% and 99% alignment identity) using metagenomic datasets spanning a range of size and diversity (Suppl. Table 1). Runtime (core time) of Nonpareil with k-mer kernel increased approximately linearly with metagenomic dataset size (Suppl. Fig. 2). For comparison, the core runtime for the alignment kernel using the same computer was over 24 h for the Lake Lanier 2009B dataset, an increase of over 250-fold.

A.2.2. Parallelization for High-Performance Computing

Nonpareil 3 can use multi-node and multi-thread parallelization using Message Passing Interface (MPI) and pthreads, respectively. We executed Nonpareil with alignment kernel using both parallelization methods, independently and in combination, on a 2.3 Gbp test dataset in order to evaluate the speedup. Compared to Amdahl's law, Nonpareil 3 speedups correspond to around 99.5% parallelization with MPI alone and around 99.8% with MPI and four threads (Suppl. Fig. 3). This means that it is possible to reduce running times as much as 200-500 times, and possibly more with additional threads per node. The observed speedup for multi-threads in one node was nearly linear.

A.2.3. Error correction testing

Low and high coverage simulated metagenome datasets of 101 bp reads (507,813 and 7,093,697 reads, respectively) were generated from 30 bacterial and archaeal genomes using a previously described method ((Rodriguez-R &

Konstantinidis, 2014c); Suppl. Table 2). For each test, 10,000 reads were randomly selected from the simulated datasets and modified with substitution errors at a constant error probability per base. The 10,000 reads were then used in Nonpareil-k as query sequences with error correction enabled and with error correction disabled (Suppl. Table 3; Suppl. Fig 4). In general, error rates affected the estimation of both coverage and required sequencing effort when error correction was disabled, but not when enabled, indicating that the correction effectively removes the effect of sequencing errors when the error probability estimation is accurate.

Supplementary Table 1. Kernel comparison of Nonpareil estimates for publicly available datasets. The k-mer kernel (K) was compared to the alignment kernel (A) with identity thresholds of 95% ($A_{95\%}$) and 99% ($A_{99\%}$). The comparisons include CPU time (*estimated for the alignment kernel in Iowa Soil, observed for all the other values), estimated abundance-weighted average coverage, and projected required sequencing effort to reach 95% coverage in samples from HMP (posterior fornix, tongue, stool), freshwater (LL: Lake Lanier), and soil (Iowa continuous corn field).

Sample	Identifier	Reference	Size (Gbp)	CPU Time (m)		Coverage (%)			Req. effort (Gbp)		
				A	K	$A_{95\%}$	$A_{99\%}$	K	$A_{95\%}$	$A_{99\%}$	K
P. fornix	SRS063417	(Human Microbiome Project Consortium, 2012)	0.01	15.7	0.08	88	79	82	0.028	0.033	0.027
Tongue	SRS055495		0.22	286	0.68	67	53	59	1.09	1.87	1.21
Stool	SRS015540		0.32	438	0.85	79	66	69	1.04	2.00	1.39
LL 2011	SRR948155	(Rodriguez-R & Konstantinidis, 2014c)	2.95	4,397	16.5	83	73	77	6.78	14.4	8.79
LL 2009A	SRR096386	(Oh et al., 2011)	1.17	1,444	6.40	61	56	61	6.46	9.13	6.18
LL 2009B	SRR096387		1.12	1,463	5.75	69	56	61	4.97	8.03	5.36
Iowa soil	JGI 402461	Not published	14.5	22,806*	49.0	53	41	44	137	172	149

Supplementary Table 2. Genomes used for simulated datasets to evaluate error correction.

Entry	Size (bp)	Abundance (%)[1]
NC_015189.1	5,178	6.33
NC_010940.1	3,533	1.00
NC_010580.1	181,736	1.78
NC_017867.1	1,938,822	2.71
NC_015052.1	2,400,312	0.45
NC_015920.1	29,448	1.02
NC_009713.1	3,678	4.58
NC_015914.1	6,221,273	4.95
NC_008255.1	4,433,218	0.06
NC_011661.1	1,855,560	0.54
NC_017910.1	4,158,725	2.20
NC_017631.1	5,131,397	5.55
NC_011419.1	100,021	1.27
NC_017642.1	90,868	1.75
NC_014558.2	53560	10.8
NC_013199.1	2968598	5.25
NC_008576.1	4719581	1.51
NC_014002.1	2012424	1.57
NC_010727.1	25164	0.54
NC_014563.1	165693	7.18
NC_016818.1	4861101	2.56
NC_007908.1	4712337	2.84
NC_015729.1	63532	0.90
NC_007954.1	4545906	4.43
NC_010506.1	5935403	5.19
NC_014009.1	5398	10.0
NC_017098.1	3285855	1.00
NC_015757.1	3551206	4.74
NC_017268.1	1139281	1.69
NC_016842.1	1139330	5.58

[1]Relative abundance of reads.

Supplementary Table 3. Nonpareil error correction for k-mer kernel.

Error Rate (%)	Coverage (%)			Req. effort (Mbp)		
	Expected	On[1]	Off[2]	Expected	On[1]	Off[2]
0.0%	47.59%	47%	47%	316[3]/254[4]	260	260
0.1%	47.59%	47%	46%	316[3]/254[4]	260	270
1.0%	47.59%	47%	37%	316[3]/254[4]	260	370
2.0%	47.59%	47%	33%	316[3]/254[4]	260	460
0.0%	98.01%	98%	98%	-[5]	-[5]	-[5]
0.1%	98.01%	98%	96%	-[5]	-[5]	-[5]
1.0%	98.01%	99%	78%	-[5]	-[5]	1,220
2.0%	98.01%	99%	61%	-[5]	-[5]	3,180

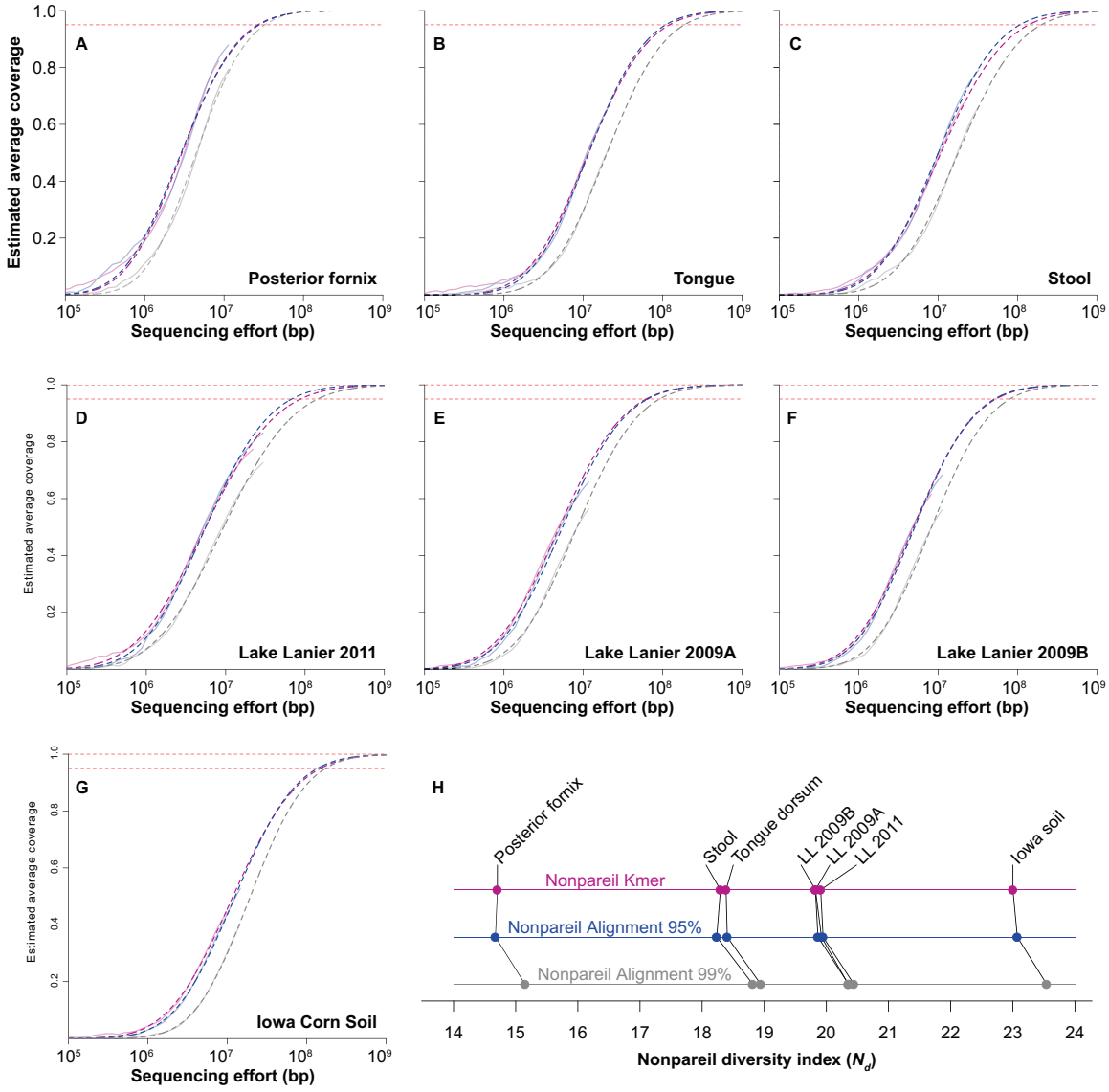
[1]Nonpareil with k-mer kernel and error correction.

[2]Nonpareil with k-mer kernel and without error correction.

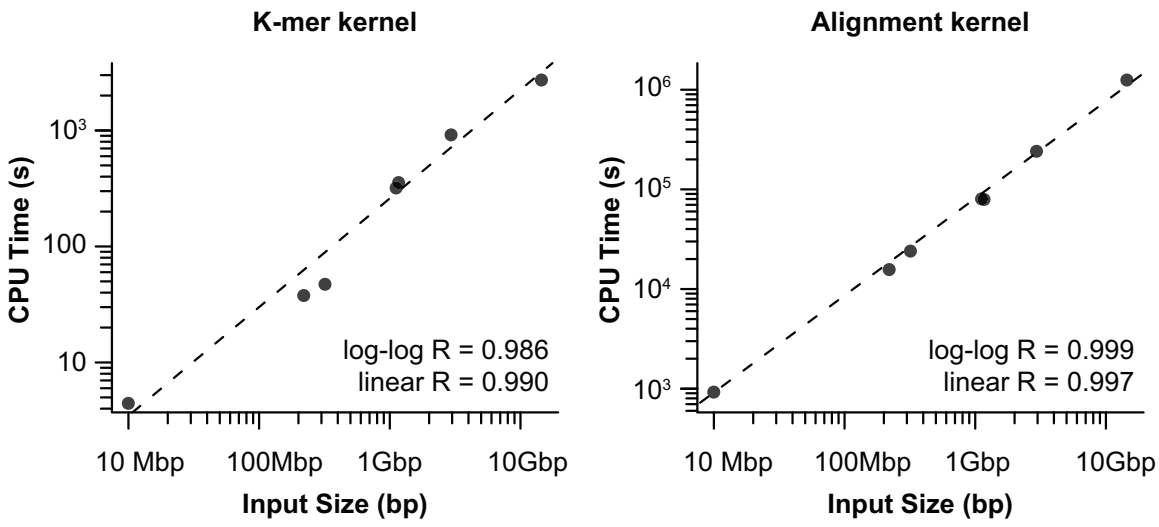
[3]Based on direct estimation in random subsampling.

[4]Based on Nonpareil with alignment kernel and no error (post-trimming).

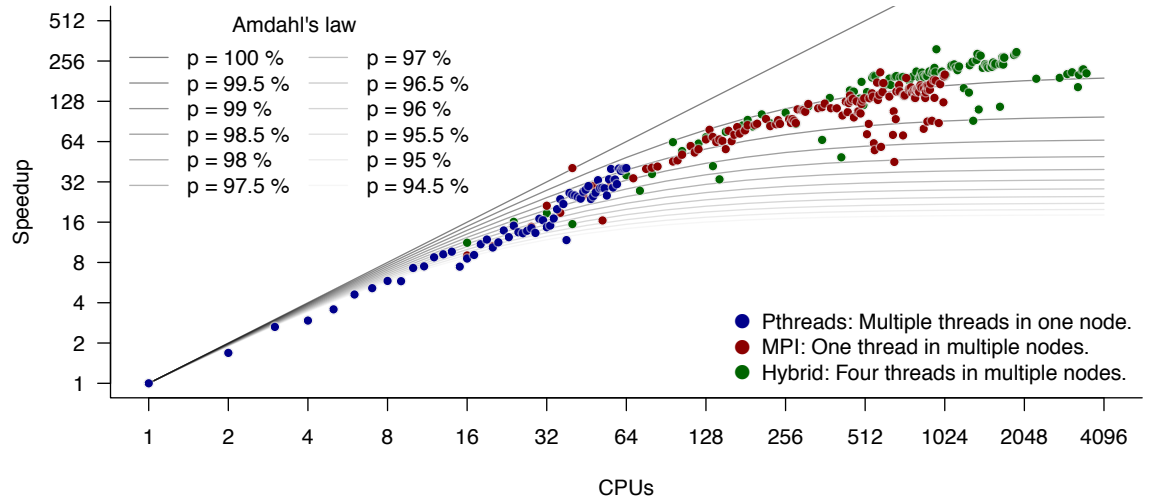
[5]Not estimated for datasets with coverage above 95%.



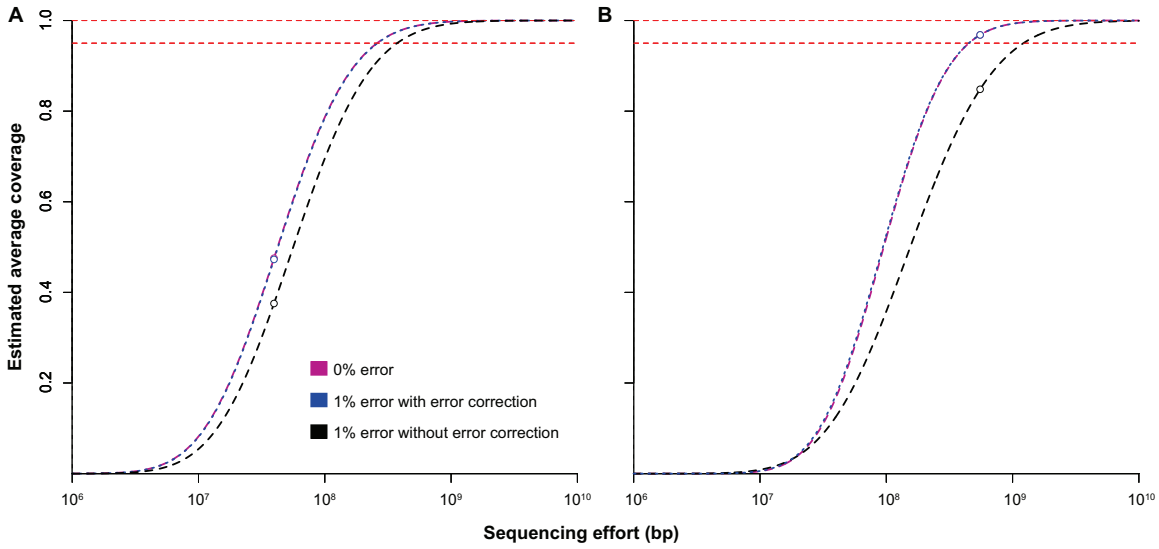
Supplementary Figure 1. Nonpareil curves for datasets highlighted in this study using k-mer and alignment kernels. A-G. The Nonpareil curves show the fit of coverage per sequencing effort to a sigmoidal model. The lines indicate coverage estimates from subsamples (solid) and Nonpareil projection curves (dashed). Horizontal red dashed lines indicate 95% and 99% coverage. Colors indicate the kernel and parameters used: Kmer kernel in magenta, alignment kernel with 95% identity in blue, and alignment kernel with 99% identity in grey. **H.** Nonpareil diversity index (N_d) for the above curves.



Supplementary Figure 2. Linear complexity of Nonpareil with k-mer and alignment kernels. Dataset size and running times from Suppl. Table 1 showing a linear time increase with input size.



Supplementary Figure 3. Speedup per processors in Nonpareil. Nonpareil estimations of the same dataset (LL_1007B; (Rodriguez-R & Konstantinidis, 2014c)) were performed with alignment kernel and default parameters in multiple processors of a node (blue), a single processor of multiple nodes (red), and 4 processors in multiple nodes (green). The base time (one processor in one node) was 82.98 h.



Supplementary Figure 4. Nonpareil curves for k-mer kernel with and without error correction. A: low coverage (48%). B: high coverage (98%).